UNIVERSITY OF ALBERTA

THE VALIDITY OF CLINICIANS' JUDGEMENTS IN STANDARD SETTING

BY

OLIVE H. TRISKA

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

Edmonton, Alberta

Fall, 1996

National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file *Votre référence*

Our file *Notre référence*

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-612-18119-7

Canada

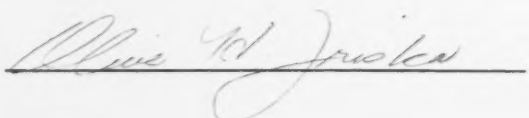
UNIVERSITY OF ALBERTA

RELEASE FORM

Name of Author:	Olive Helen Triska
Title of Thesis:	The Validity of Clinicians' Judgements in Standard Setting
Degree:	Doctor of Philosophy
Year this Degree Granted:	1996

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



12306 - 91 Street

Edmonton, Alberta

Canada, T5B 4C5

August 28, 1996

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND REASEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled THE VALIDITY OF CLINICIANS' JUDGEMENTS IN STANDARD SETTING submitted by Olive Helen Triska in partial fulfillment of the requirements for the DOCTOR OF PHILOSOPHY.

Dr. Thomas O. Maguire

Dr. W. Todd Rogers

Dr. Charles C. Harley

Dr. Ernest N. Skakun

Dr. Jennifer A. Crocket

Dr. John A. Hattie

July 3, 1996

DEDICATION

I would like to dedicate this work to

my parents

Helen Triska and the late Nicholas Triska.

my brother

Terrace Nicholas

my niece and grand-nieces

Angelene, Angelica, and Devin

ABSTRACT

A premise of this study was that to set credible standards of student performance on a particular item, clinician judges must possess a level of competence that is analogous with the declarative and procedural knowledge implied by that item. This study examined the validity of clinicians' judgements in standard setting, using a distractor approach, when placed in a cognitive psychology framework. To investigate the validity of the clinician's judgements, three theories/models were used: Anderson's (1983) information processing theory (ACT*); (2) Royer, Cisero, and Carlos' (1993) assessment of cognitive skills model; and, (3) the Nedelsky (1954) approach was used to set standards.

The 12 clinicians who participated in this study performed four specific tasks while verbalizing their thoughts for 10 multiple-choice items. To explore the data, a multilayered-multifaceted qualitative and quantitative analysis was used. Student protocols taken from Skakun (1994) were compared with the clinicians' perceptions of student reasoning.

A preliminary analyses suggested that the selection of standard setters be chosen according to their level of domain competence and whether they are expert or novice reasoners. Differences were present between the passing scores between the clinicians who chose the keyed response and those clinicians who did not choose the keyed response. The competently reasoning clinicians set substantially higher passing scores and made more consistent decisions than the inaccurately reasoning clinicians. Competently reasoning clinicians, after being categorized as experts or novices, showed differences in their judgements for standard setting. There appeared to be consistency within each group, but little consistency between the groups.

The validity of the Nedelsky Type (NT) judgements were investigated by using an alternate distractor based method, a competency rating scale (CRS). When the strength

of relationship between the NT and CRS judgements was calculated, evidence supported the validity of the NT approach. Using the Chi-Square test with Yates' correction as a standard for evidence in support of the validity of the NT approach, it was shown that the clinicians' judgements were significantly comparable for eight items using the CRS approach. In the two remaining items, the discrepant judgements were made by the expert clinicians.

To validate the clinicians' perceptions of how students eliminated distractors and reason to choose keyed response, four items were examined from two specific clinician perspectives. First, the clinicians' justifications for elimination of distractors students should know are incorrect were examined. The results showed that the expert clinicians were more successful in identifying reasons students used to eliminate distractors as compared to the novice clinicians. The findings suggested that the novice clinicians may have moved down the knowledge continuum to the students position more successfully than expert clinicians.

Ten findings with implications and suggestions for further research for each finding are presented. Three recommendations for standard setting are offered.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation and special gratitude to my advisor, Dr. Thomas O. Maguire for his patience and thoughtful consideration of my views on competency, measurement, and my manuscripts. His suggestions for improvement continually provided me with insights into the nature of behavioral research, clear thinking, and accurate expression.

Dr. Todd W. Rogers generously gave his time to serve on my candidacy and dissertation committees. It was through his expertise and guidance that I was able to complete this massive study.

Dr. Charles C. Harley, my medical advisor, brought several thoughtful perspectives to this study that would have been otherwise overlooked. His understanding of medical education and careful clarification issues was invaluable.

Dr. Ernest N. Skakun gave me liberal amounts of time to explain various germane issues concerning competency and reasoning. He generously shared the students' protocol with me to validate the clinicians' perceptions of student problem solving strategies. Without Ernie's support, this study would have been severely compromised.

I would like to express my gratitude to Dr. Jennifer A. Crocket and Dr. John A Hattie for serving on my committee and providing insights into this study. Their comments were both thoughtful and thought provoking.

I would like to acknowledge faculty members who have contributed substantially to my education. In particular, I am indebted to Dr. Todd W. Rogers, who encouraged me to return for doctoral study. Dr. Steven Hunka's wise words of support and reassurance during my program were most helpful and appreciated. I also want to thank Lana and Rena whose smiles and support helped guide me through many practical

aspects of graduate school.

No research project of this magnitude could be completed without the support and encouragement of other significant individuals in my life, both far and near. I would like to express my appreciation to my mother for her sustained interest and encouragement. My brother, niece, and grand-nieces provided diversion when it was needed, just to keep everything in perspective. My heartfelt thanks goes to Norman "Muggsy" Forbes, who kept me afloat with his special brand of humour during the final stages of my work. To all of you, thank-you for your constant interest and support!

This research was supported in part by a National Health Research and Development PhD Fellowship, a Walter H. Johns Graduate Fellowship, and a University of Alberta PhD Scholarship.

TABLE OF CONTENTS

Chapter 1 - Purpose of the Study 1
Introduction 1
Background 2
Overview of Dissertation 5

Chapter 2 - Related Literature, Theory, and Models 7
Introduction 7
The Literature 8
Standard Setting 8
Cognitive Psychology 20
Theoretical Models Used in This Study 24
The ACT* Theory of Cognitive Skill Acquisition 25
The Cognitive Skill Assessment Model 29
The Nedelsky Procedure for Standard Setting 33
Summary 34

Chapter 3 - Method and Preliminary Analysis 36
Introduction 36
Relevance of Preliminary Analysis 36
Selection of the Clinicians 37
Characteristics of the Students 38
Rationale for Using the Students Protocols From Skakun's Study 38
Characteristics of the Students 38
Description of the Multiple-choice items 38
Data 40
Preliminary Analysis 41
Competency 41
Assessment of Reasoning 42
Verification 42
Results 45
Summary 48
Implications of the Results on the Judgement Comparisons 49

Chapter 4 - Judgement Comparisons 50
Introduction 50
Rationale for Comparisons of Clinicians' Judgements 50
Rationale for Using an Alternative Distractor Based Standard Setting Approach 51
The Clinicians' Judgements Using an NT Distractor Based Procedure 51
Comparison of the NT Judgements Between G1 and G2 53
Comparison of the NT Judgements Between G1E and G1N Clinicians 56
Comparison of Clinicians' Judgements Using the NT and CRS 58
Method 59
Results of the NT and CRS Comparison 62
Summary of the Judgement Comparisons 68
Implications of the Clinicians' Perceptions of Competent Reasoning in Students 70

Chapter 5: Clinicians' Perceptions of Students' Reasoning 71
Introduction 71
Rationale for Analyzing the Clinicians' Perceptions
of the Students' Reasoning 71
Method 72
Distractor Analysis 72
Reasoning to the Keyed Response 74
Overview of the Chapter 74
The Clinicians' Perceptions of the Students' Competent Reasoning 75
Item 347 75
Elimination of Distractors 75
Reasoning to the Keyed Response 78
Item 317 81
Elimination of Distractors 81
Reasoning to the Keyed Response 84
Item 733 87
Elimination of Distractors 87
Reasoning to the Keyed Response 90
Item 332 93
Elimination of Distractors 93
Reasoning to the Keyed Response 96
Summary of the Clinicians' and Students' Reasoning Comparisons 100

Chapter 6 - The Findings, Implications, and Future Research 104
Introduction 104
Organization of the Chapter 104
Standard Setting Using the Nedelsky Type Approach 104
Standard Setting Within a Cognitive Psychology Framework 108
Measurement Issues Related to Standard Setting 110
Recommendations for Standard Setting Procedures 112
Closing Remarks 113

References 114

Appendix A: Multiple Choice Item Used in the Study 120

Appendix B: Clinicians' Interview Package 122

Appendix C: Contingency Tables of the Clinicians' Decisions Comparisons 130

Appendix D: Distractor Verification Analysis 135

Appendix E: Clinicians' Perceptions Verification Analysis 136

LIST OF TABLES

Table 3.01
Summary of the Clinicians' Responses 42

Table 3.02
Description of Cognitive Assessment of Experts 43

Table 3.03
Description of Cognitive Assessment of Novices 44

Table 3.04
Results of the Cognitive Assessment 45

Table 3.05
Z-test to Determine the Difference Between the G1E Clinicians and Expert Students 46

Table 3.06
Frequency of G1 Clinicians' Judgements 47

Table 4.01
Passing Scores Set by G1 and G2 54

Table 4.02
Agreement of the NT Judgements Within and Between G1 and G2 Clinicians 55

Table 4.03
Passing Scores Set by G1E and G1N 56

Table 4.04
Summary of the NT Judgements Between G1E and G1N Clinicians 57

Table 4.05
The Competency Rating Scale 59

Table 4.06
2 X 2 Contingency Table for Item 317 for G1 60

Table 4.07
Phi Coefficient and Chi-Square Test with Yates' Correction:
NT and CRS Judgements 61

Table 4.08
2 X 2 Contingency Table for Item 267 for G1 63

Table 4.09
2 X 2 Contingency Table for Item 267 for G1E 63

Table 4.10
2 X 2 Contingency Table for Item 267 for G1N 63

Table 4.11
2 X 2 Contingency Table for Item 771 for G1 64

Table 4.12
2 X 2 Contingency Table for Item 771 for G1E 65

Table 4.13
2 X 2 Contingency Table for Item 733 for G1 65

Table 4.14
2 X 2 Contingency Table for Item 733 for G1E 66

Table 4.15
2 X 2 Contingency Table for Item 733 for G1N 67

Table 5.01
Item 347 Summary of Reasons for Eliminating Distractors 77

Table 5.02
Clinicians' Perceptions of and Students' Behaviours for Item 347 79

Table 5.03
Item 317 Summary of Reasons for Eliminating Distractors 82

Table 5.04
Clinicians' Perceptions of and Students' Behaviours for Item 317 84

Table 5.05
Item 733 Summary of Reasons for Eliminating Distractors 88

Table 5.06
Clinicians' Perceptions of and Students' Behaviours for Item 733 90

Table 5.07
Item 332 Summary of Reasons for Eliminating Distractors 94

Table 5.08
Clinicians' Perceptions of and Students' Behaviours for Item 332 96

Table 5.09
Summary of the Proportion of Agreement of Reasons for
Eliminating Alternatives Between the Clinicians and Students 100

Table 5.10
Summary of the Clinicians' Perceptions of the Students' Reasoning to Solve Items .. 102

LIST OF FIGURES

Figure 01. Continuum of Medical Education and the ACT* Model. 28

Figure 02. The ACT* Model, Assessment of Reasoning, and Medical Education. . . . 32

Chapter 1 - Purpose of the Study

Introduction

The notion of professionals being evaluated to determine their level of competency before starting their professional practice is not new. This study focussed on one specific type of professional competence, medical practice. To investigate the issue of professional competence in this area, the research question that prompted this study was: "How valid are judgements in standard setting, using a distractor approach, when placed in a cognitive psychology framework?"

Distinguishing between professional competence and incompetence is essentially a judgemental process. Although there are strong empirical components in the final analysis, the ways in which the decisions are made are usually based on informed judgements. Previous literature has dwelt on the technical aspects of competence, evaluating instrument construction, and standard setting. But, a logical prior consideration is the determination of what constitutes competence. This requires not only deciding on the combination of skills and knowledge necessary to carry out professional tasks, but also estimating the level of competence required for professional practice.

Competence is largely covert. Although there are demonstrable clinical skills required to be a successful physician, much of competence refers to mental processes and attitudes. Generally, individuals who aspire to achieve professional status progress along a continuum of professional training, are evaluated, and are then deemed certified and licensed as practitioners. Certified individuals work autonomously and collaboratively within the practices unique to their professions.

The psychological constructs used by experts or judges to determine the level of competency required to practice professionally are only implied in empirical studies reported in the literature. The concept of competence appears to be an individual psychological notion constructed in the minds of judges. Tacit knowledge appears to be a critical component, as is experience, memory, speed of decision-making, and certain psychomotor skills.

Certification generally relies on evaluation of competence. Certification tests determine levels of competence in one's professional practice with two critical categories: those individuals who meet the quality of the standard and those who do not. Those who score above the passing point are certified to practice medicine, and those who fall below this point are not. The setting of passing points is an important activity, but it was not within the scope of this study. The present study focussed on the judges' psychological constructs that formed their decisions while setting standards.

Background

The purpose of this study was to gain insight into the complex issue of the validity of clinicians' judgements while setting standards. Before individuals set standards of competence, however, a the definition of competence is required.

The definition of competence, and its role as a psychological construct, has been investigated by several researchers (Chi, Glaser, & Rees, 1981; Doolittle & Yekovich, 1994; Groen & Patel, 1989; Kane, 1992; LaDuca, 1980; McGaghie, 1980; Norcini, Lipner, Langdon, & Strecher, 1987; and, Patel & Groen, 1991a). More specifically, two perspectives have been investigated, the performance of experts' and the experts' cognitive processes. Experts and nonexperts in various professions were observed during their professional activities by Ayton (1992), Chi et al. (1981), Doolittle and Yekovich (1994),

Groen and Patel (1989), Lesgold (1983), Patel and Groen (1991b), and Shanteau (1992). These researchers used quantitative methods to show that when experts and novices solve problems, they differ in: (1) the speed of information processing; (2) memory span; and, (3) the types of strategies used to solve problems. There seemed to be a covert cognitive reasoning component that was enacted within both experts and novices. Assessment of this "component" was difficult because of the researchers' inability to identify these variables (Chi, Glaser, & Rees, 1981; Johnson, 1989; Kane, 1992; Keren, 1992; LaDuca, 1980; McGaghie, 1991; McGaghie, 1993; Meskauskas, 1986). One result of the "component" dilemma is that there is no agree on the cognitive variables that underlie competence. Because this qualitative component has not been clearly identified, measurement tools to assess this "component" are meagre.

Judges have difficulty agreeing upon standards of practice acceptable to the profession and society. They have difficulty verbalizing and agreeing on competence because each judge has an individual notion of competence. There are examples of operational definitions of competence in the literature, but judges cannot decide on which external behaviours are manifestations of competent reasoning.

To link clinical reasoning problem solving skills with problem solving behaviour, information processing theory and assessment of cognitive skills have been used. Various theoreticians have proposed theories on how information is processed, but at issue is having a theory consistent with the continuum of medical education knowledge and skill acquisition. Anderson's ACT* (Active Control of Thought) theory, (Anderson, 1983; Anderson, 1993) of the architecture of cognition in combination with the assessment of cognitive skills (Glaser, Lesgold, & Lajoie, 1985) may provide the crucial bridge in assessing competent clinical reasoning skills and behaviours.

Royer, Cicero, and Carlo (1993) implied that the ACT* theory and assessment of cognitive skills could be placed on similar continua. By integrating these two models with the continuum of development clinical reasoning in medical students, Doolittle and Yekovich (1994) explained that clinicians could probably assess the level of clinical reasoning skills that indicate competent problem solving behaviours. They gave an example of the process of knowledge acquisition in surgery by charting the development of knowledge from a first year medical student to a surgical resident. Their explanation illustrated the progression of knowledge accumulation according to the ACT* theory (Anderson, 1983, 1993) in concert with assessment of cognitive ability as explained by Glaser et al. (1985).

Skakun, Maguire, and Cook (1994) reported that students tended to solve items using a variety of behaviours, a total of 25. Students were inclined to solve items by attending to the stem (five strategies), then examine each alternative in order to match the information with the scenario stated in the stem (16 strategies). These researchers explained that successful problem solvers used one of four different overarching strategies to reason to their answer. Skakun et al showed that when students solve items, they used discrete bits of data, facts, and definitions, and elaborated on these chunks of information.

As explained earlier, the focus of standard setting has been on technical aspects and not on the judges' constructs upon which the validity of their decisions were based. In order to determine if standards of competence are congruent with the students' cognitive structures, judges should reason at a similar level as the students, and possibly use similar strategies to solve the item. This implies that standards based on how students think, and not on reasoning skills of hypothetical borderline candidates that may indicate competent reasoning.

Nedelsky (1954) proposed a distractor based approach to setting passing scores on multiple-choice tests. He did not describe a theoretical base for making standard setting decisions, but in this study a theory that supports his technique is described in which each distractor, in addition to the keyed response, is dealt with individually by each judge. To focus this study, I imposed the following three constraints.

1. The specific area of investigation was the validity of clinicians' judgements in standard setting when placed in a cognitive psychology framework.
2. The foundation this research rested on was Anderson's ACT* theory (Anderson, 1983; Anderson, 1993) of skill acquisition, and Glaser's (Glaser et al., 1985) taxonomy assessment of cognitive skills, and a distractor based approach to setting standards, the Nedelsky (1954) procedure.
3. This study was limited to using the multiple-choice item format only.

Overview of Dissertation

This study was designed to investigate the validity of clinician's judgements in standard setting from a cognitive psychology perspective. Chapter 1 explained the issue of competence in relation to setting standards, the purpose of this investigation, and the delimitations imposed on this study by this researcher. Chapter 2 is divided into two major sections, the literature relevant to this study, and the theoretical tenets this study upon which it rests.

Chapter 3 presents the relevance of a preliminary analysis, then describes the population, the multiple-choice items used in the study, and outlines the data collection procedures. This chapter ends with the results of the preliminary analysis.

Chapter 4 is divided into five sections. First, the rationales for comparisons of clinicians' judgements and for using an alternative distractor type standard setting process

are explained. Second, the analyses and results of the standard setting procedure for elimination of distractors are presented for the clinicians. Third, comparisons of judgements made within the competently reasoning clinicians are described, and the implications for the remainder of the study are presented. Fourth, the analyses and the results of the comparison of judgements between the standard setting procedure judgements and the alternative distractor approach within the competent clinicians are described. Finally, a summary of the results and the implications for a theory of standard setting are presented.

Like Chapter 4, Chapter 5 is organized into sections. In section one, the rationale for the analysis and results of the clinicians' and students' protocols is explained. Section two consists of a discussion about how the clinicians set standards by eliminating alternatives and explain the misconceptions in each alternative. Then, clinicians explained how competent students reasoned to select the keyed response in section three. Student protocols taken from Skakun (1994) were compared with the clinicians' perceptions of student reasoning in sections two and three. Chapter 5 ends with a discussion that integrates the results of these analyses.

Chapter 6 focus on a discussion of the preliminary analyses, the judgement comparisons, and an interpretation of reasoning comparisons and relevant literature. Chapter 6 closes with suggestions for future research.

Chapter 2 - Related Literature, Theory, and Models

Introduction

The setting of performance standards has occupied psychometricians' research agendas for four decades (Nedelsky, 1954), and continues to the present generation of researchers (Cizek, 1991; Mills, 1995). Various standard setting procedures were developed to reflect diverse ideologies and for professional credentialling and licensing bodies in medicine and the allied health fields (Meskauskas, 1986). Research on these procedures focussed on the technical aspects of standard setting and not on the underlying constructs of the judges' decisions that drive the judgement process. Meskauskas pointed out that the cognitive processes of the judges needed exploring, but offered no suggestions on how to approach this exercise. To gain insight into the cognitive processes facing judges when they determine performance standards, this study addressed the question, "How valid are clinicians' decisions in standard setting when placed in a cognitive psychology framework?"

To investigate this concern, the discussion in Chapter 2 is divided into two major sections, the literature relevant to this study, and the theoretical tenets this study rests upon. In section one an overview of literature pertinent to setting standards in education and medical education will be presented. Then, relevant research in cognitive psychology and cognition in medical education will be addressed. In section two the theoretical notions regarding cognitive psychology, assessment of reasoning skills, and the standard setting procedure used in this study are presented. Each segment within the major sections is addressed, beginning with a discussion of setting standards in education and medical education.

The Literature

Standard Setting

Researchers have approached setting standards of performance from heterogeneous points of view for the past four decades. Nedelsky (1954) proposed a standard setting procedure for multiple-choice item tests using a judgemental model based on a knowledge continuum. He explained that each multiple-choice examination item should be examined individually by a group of judges. When judges examine each item's alternatives, they should be able to identify which options competently reasoning students should eliminate as incorrect. They should know which options are moderately wrong, an indication of partial knowledge, and which option is the correct response, the keyed option. The number of remaining options is totalled, and a reciprocal is calculated. The reciprocals are summed, resulting in the passing score of a "barely passing student".

Glaser's (1963) seminal article on criterion-referenced testing brought the issue of mastery to the forefront of evaluation. He explained that students ought to be tested on what they know or accomplished, not in relation to their standing in a group. New standard setting methods were developed consistent with his notion of mastery and were evaluated. In the following decade many articles appeared in the psychometric literature.

Angoff (1971) proposed an absolute standard setting procedure, approaching standard setting from a holistic demographic perspective. He asked the judges to provide an estimate of the probability that a minimally competent individual would correctly answer the item. The item probabilities are summed for the passing score for the total test. Unlike the Nedelsky approach, judges do not identify which distractors a 'minimally competent' individual would eliminate, but view the item as a whole.

Ebel (1972) developed a method to classify an item according to two dimensions, relevancy and difficulty. Four relevancy categories in conjunction with three difficulty levels formed a 4 X 3 matrix into which the judges classified all items. Judges estimated the percentage of items in each cell that a minimally competent individual should answer. The products of the number of items in each cell and the appropriate percentages are summed and divided by the total number of items, resulting in the passing score for a borderline candidate.

Andrew and Hecht (1976) did one of the first comparisons between the Nedelsky and Ebel procedures standard setting methods. One objective of their study was to determine "whether the average judgements made by individuals within each group would differ significantly from the group judgements arrived at by consensus" (p. 46). Two groups of judges, each consisting of four individuals, met separately twice to set the standard on 180 multiple-choice items that were part of a nationally administered certifying examination in the health professions. Using a split-half odd-even method of comparisons between the items, with one group using the even numbered items, and one group using the odd numbered items, they confirmed their assumption of equivalence between the two forms. They found that the judges individual decisions did not differ significantly from the consensus judgements of the total group. Analysis of variance using group judgements as the dependent variable and groups (Factor A) and procedures (Factor B) as the independent variables showed that the F ratio was significant beyond 0.01 for the two standard setting procedures. No significant interaction effect (0.05 level) between Factor A and Factor B was observed. These results implied three things: that the Nedelsky and Ebel methods for setting standards resulted in different standards for parallel samples of test content, different groups of judges set similar standards using the same examination content, and the individual judgements were not significantly different

from the group judgements. However, Andrew and Hecht failed to examine the intrajudge psychological constructs used to make the decisions were not included in this seminal study.

In 1976, Meskauskas summarized the standard setting models and classified them into two broad categories: continuum models in which mastery was viewed as an area on a continuum, and state models in which mastery was either present or absent. He went on to review three continuum models, Nedelsky (1954), Ebel (1972), and the Kriewall (1972) binomial-based model. Meskauskas then reviewed two state models, Emrick's (1971) mastery testing model and Roudabush's (1974) dichotomous true-score model, and two decision methods not referenced to mastery models, Millman's (1972, 1973) binomial-based model and the Davis and Diamond (1974) Baysean method. Lastly, Meskauskas commented on the work of Novick and his collaborators (Hambleton & Novick, 1973; Novick, Lewis, & Jackson, 1973; and Novick & Lewis, 1974). In summarizing his research findings, Meskauskas pointed out that: (1) standard setting methods need validating; (2) reliable ways of obtaining information to use the standard setting model are needed; and (3) predictions generated by the models need validating. Meskauskas concluded his paper by advocating a team approach to evaluation to understand the application of both theory and practice. Following Meskauskas's paper, the *Journal of Educational Measurement* devoted an entire issue in June, 1978 to standard setting.

Research in which psychometric properties of various methods were compared (Brennan & Lockwood, 1980; Cross, Impara, Frary, & Jaeger, 1984; Koffler, 1980; Mills, 1983) continued into the 1980's. A summary of this literature is found in Jaeger (1989). These researchers provided insights into advantages and disadvantages associated with the various methods. Throughout these investigations these researchers advocated the Angoff procedure and its modifications became the method of choice. The investigation of the

cognitive psychological aspects of the judges, those individuals' whose decisions were scrutinized, was negligible. However, identification of what constitutes a "minimally competent" individual, the cornerstone of this method, has caused concerns among the judges.

The elusive definition of a barely qualifiable or minimally competent individual was operationally defined in the literature, but how these minimally competent individuals thought was not explored. As recently as 1993, Cizek (1993) expressed concern with the definitions of parameters for minimal competence and borderline ability-both of which are indistinct concepts. Mills (1995) explained that,

> An explicit definition of minimal competence is required for most standard setting procedures. Simply put, minimal competence is the "minimal level of knowledge and skills required for licensure." Unfortunately, this simple definition is not an operational definition of minimal competence and, therefore, is inadequate given the variety of skills being tested, the different ways they can be acquired, and the possible compensatory effects that strengths in one area might have for weaknesses in another area. (p. 241)

A tremendous interest surged among researchers to develop appropriate standard setting methods for criterion referenced tests. This wave of standard setting research in education was summarized by many researchers (Berk, 1986; Jaeger, 1989; Livingston & Zieky, 1982; Shepard, 1980, 1984). Shepard (1984) started her article by stating, "As with any other psychological construct that cannot be embodied by concrete test scores, performance standards pose special problems for measurement experts. . . . The most elaborate domain specifications in the world cannot compensate for invalid standards." (p. 169). She went onto review standard setting methods, both state and continuum mastery models, other empirical models, classification errors, and outlined guidelines for practitioners for setting standards. Shepard also acknowledged the standard setting

dilemma of the judges when determining a cutoff score on the test. Shepard linked standard setting with the intrajudges' psychological constructs by saying, "The methods are strategies to try to elicit both the absolute and normative standards *in the judges*. The standard we are groping to express is a psychological construct in the judges' mind rather than *in the methods*." (p. 188). Clearly, the underlying psychological constructs were pivotal to setting standards, but standard setting techniques were still "centre stage", instead of the judges' domain competency or level of expertise.

Berk (1986) identified and reviewed 38 methods for setting standards on criterion-referenced tests based on state or continuum models. He subdivided the continuum model into three subsections, judgemental, judgemental-empirical, and empirical-judgemental. The empirical-judgemental subsections were further divided into two groups, setting standards and adjusting standards. Berk presented elaborate tables evaluating the various models. He reviewed the technical adequacy and practicability ratings of 23 standard-setting methods. Finally, Berk offered specific guidelines for the different types of consumers who were required to choose a method of setting standards, such as classroom teachers, educational certification test specialists, licensure and certification boards, and test publishers and independent test contractors. The common tread in these guidelines was that the selection of the standard setting method should be based on the use of the judgements, performance data, or a combination of the two. Berk advocated the use of the judgemental-empirical model for certification and licensure for the professions and occupations, that is, some variation of the Angoff (1971) method. Throughout this treatise on setting standards, little attention was paid to the intrajudge psychological constructs on which the judgements were made.

Research on procedures for standard setting was examined by various psychometricians. Sources of inconsistency and variability within the standard setting

process were first acknowledged by Nedelsky (1954). Early in his paper, Nedelsky linked the instructor's judgements with students' cognitive development, and cautioned his readers on the consequences of the instructor's decisions. Nedelsky alerted the reader to this issue by stating:

> Judging a response in comparison with other responses is theoretically sound, for it probably more closely corresponds to the mental processes of the student. To make a proper judgment of this kind, requires time and considerable pedagogical and test-wise sophistication; with responses more heterogeneous than in the example cited a reliable judgement may be impossible." (p. 7).

Nedelsky implied that judges needed: sufficient time to make their judgements, to be knowledgeable in their field, and to be trained in objective item testing.

Hambleton and Powell (1983) offered guidelines for helping standard setting committees or groups addressing issues on technical matters. They suggested a series of issues or guidelines to consider when selecting the judges, judging items, the nature of the judgemental process, use of other information, and data analysis. Giving judges ample time to make their decisions was absent from their suggestions. Other researchers also did not acknowledge this factor in setting standards (Berk, 1986; Busch & Jaeger, 1990; Mills, 1995). They also discussed the standard setting process, but providing the judges enough time to deliberate was assumed. Tying the validity of the judges decisions with time needed to make those decisions was absent in the standard setting literature.

Individual's partaking in standard setting activities are thought to be competent and expert in their areas. Hambleton and Powell (1983), when listing the criteria for judges, explained that when selecting judges their occupation, specialty, and level of educations should be considered. No advice was given how assess or determine these criteria. Plake, Melican, and Mills (1991) listed five strategies to strengthen interjudge

and intrajudge judgements, but nothing was stated about the knowledge level of the judges, only that they are made by experts. Again, nothing was suggested on how to determine expertise among the judges. Mills (1995) recognized different levels of domain expertise by saying,

> Virtually all standard setting methods require input from experts. Not all members of a profession will be qualified raters and different methods may require experts with different experience. Experts will need specific knowledge, skills, and experiences for the tasks they are to perform (p. 239).

Researchers conceded that an individual must be knowledgeable in her/his field, but offered no way to assess if a judge possesses the knowledge necessary to make valid judgements.

Hambleton and Eignor (1980), Shepard (1980), and Livingston and Zieky (1982) advocated the training of judges to reduce the variation and invalidity of the judges' decisions. Hambleton and Powell (1983) outlined a broad framework for setting standards in which they addressed the role of the judges. In their study they devoted a section to selecting and implementing a standard setting procedure. Two areas they highlighted were the judges and the nature of the judgemental process. Hambleton and Powell pinpointed 18 considerations in the two areas, but nothing was said about the judges' psychological constructs that drives their decisions in standard setting.

When Fitzpatrick (1989) investigated the social psychology on the effects of group discussion, she said that group dynamics played an important role in the decision making process. She explained that, "We noted that publicly binding people to positions they hold may make them resistant to social comparison and to changing those positions." (p. 324).

Plake et al. (1991) reviewed four sources of intrajudge inconsistencies, variables

within each judge, the standard setting process itself, validity of the judgements, and derivation of the passing score. They suggested that intrajudge consistency could be increased by providing judges with periodic retraining and group discussions, and providing data on the item performance. Plake, et al., suggested that group discussions be part of the standard setting process.

Busch and Jaeger (1990) offered insights into variables that affect intrajudges' decisions. They stated, "Although many standard-setting studies have summarized judge's characteristics, they have not examined the judges' characteristics and the standards they recommend" (p. 146). They went on to explain, ". . . we know that judges who participate in standard-setting studies are far from omniscient and vary widely in their knowledge of competency requirements and their abilities to propose reasonable and defensible test standards" (p. 147). The remainder of the study was devoted to an empirical investigation of judgements using the Angoff procedure for seven subtests of the National Teacher Examinations Communication Skills and General Knowledge Tests for 236 expert judges. No mention was made of the judges' grounds for making their judgements.

Kane (1994) presented a cogent argument on the effect of the validity of clinicians' decisions on passing scores, and provided a framework for examining the validity of performance standards for high-stakes achievement tests. In this framework, Kane emphasized "conceptual issues and broadly defined methodological questions, on the kinds of data that can be collected and on the advantages and limitations of different types of evidence" (p. 425). In the section devoted to procedural evidence for validity, he cited two reasons to validate performance standards. The first reason dealt with empirical checks of the interpretation of passing scores. Kane linked the second reason to evaluating policy decisions by saying,

> [W]e can have some confidence in standards if they have been set in a reasonable way (e.g., by vote or by consensus), by persons who are knowledgeable about the purpose for which the standards are being set, who understand the process they are using, who are considered unbiased, and so forth. (p. 437)

Further in the article, Kane (1944) stated five specific issues that directly affect the validity of the judgements in standard setting procedures. But, the reasoning judges use to make their decisions was not acknowledged in the entire document. No criteria were given on how to assess the knowledge of the individuals setting the standard. These two points, the reasoning of the judges and the knowledge level of the judges seems integral to validity of standard setting judgements, but were entirely overlooked by Kane.

Mills (1995) provided a current overview of the process of setting standards. He offered some suggestions on improving the decisions. Mill said that standard setting needed participation from experts and acknowledged that,

> Not all members of a profession will be qualified to be raters and different methods require experts with different experience. Experts will need specific knowledge, skills and experiences for the tasks they are to perform. The selection process should ensure, to the extent possible, that experts represent the full diversity of the profession and the various constituencies affected by the test. (p. 239)

Cizek (1993) contends that ". . . The theoretical underpinnings of standard setting practice have not been fully investigated and explicated. The missing framework is . . . a skeleton in the psychometrician's closet" (p. 97).

Researchers in medical education were confronting similar issues in standard setting. Skakun and Kling (1980) between the Nedelsky and Ebel procedures. Harasym (1981) investigated whether the Nedelsky and modified Angoff, were equivalent and the consequences of choosing either procedure. Harasym concluded that: (1) the Nedelsky procedure consistently produced a lower minimum passing level than the modified Angoff

procedure; (2) the lower passing level classified more students satisfactory than the modified Angoff procedure; and (3) the lower passing level produced by the Nedelsky procedure was caused by the first large increment in the passing level (e.g., 1.0 to 0.5), then smaller increments (e.g., 0.33, 0.25, 0.20) whereas in the modified Angoff, this does not occur. Harasym did not acknowledge the judgemental process judges engaged in to arrive at their decisions, but chose to concentrate on comparing the psychometric properties of both standard setting methods.

In 1986, Meskauskas presented an update of the developments in setting standard for credentialling examinations. He noticed that much attention was paid to the techniques, but setting standards involved much more. Meskauskas reviewed the developments concerning conceptual and theoretical developments from 1954 to 1985. Then, he briefly looked at the stability of passing scores associated with standard setting. This was followed by a section on five implications for the health professions. In the conclusions, Meskauskas identified two key research issues. "The first is a need to explore the determinants of intrajudge and interjudge variance in depth. . . . Second, there is a need to study further the cognitive context of the judging task" (p. 200).

As with measurement specialists in education, the Angoff method for setting standards remained a popular method of setting standards in medical education. Angoff's procedure for setting standards was used because it was understandable to judges and test users, judges became cognizant of the test items, is test centred, and it is based on absolute standard (Meskauskas, 1986). Critical to the Angoff procedure was the minimally competent individual, a great concern in the practice of medicine. Operational definitions of this hypothetical "statistical abstraction" was found in the literature. Skakun and Kling (1980) explained,

> To help judges with the conceptualization of a "barely qualifiable" candidate, judges were asked to think of candidates in their own training program who barely passed the written examinations in previous years and to use that performance of these candidates as a basis for their decisions. (p. 231)

Norcini et al. (1987) described minimally competent individuals as ". . . clinicians whose knowledge and skills, as measured by the examination, would represent the boundary between those qualified and those unqualified to receive a certificate" (p. 58). Mills (1995) approached the minimum competency controversy from a different perspective. He stated,

> Simply, minimal competence is the "minimum level of knowledge and skills required for licensure." Unfortunately, this simple definition is not an operational definition of minimal competence and, therefore, is inadequate given the variety of skills being tested, the different ways they are acquired, and the possible compensatory effects that strengths in one area might have for the weaknesses in another area. (p. 241)

Throughout the literature, operational definitions of minimal competency were given, usually from a demographic perspective, consistent with the Angoff procedure for setting standards. The implication of these definitions was that some population in a group ought not be licensed or certified. From a cognitive psychology perspective, however, competence refers to an individual's reasoning processes.

In June 1994, a set of articles on the validation process for licensing and certification test score decisions was published in *Evaluation & The Health Professions*. Various aspects of standard setting were addressed. In particular, Haladyna (1994a) confronted the role of behaviourism and cognitive psychology in psychometrics. He approached defining a construct from two perspectives, an operational definition and construct validity. He argued that, "The problem with operationism is getting two

independent test developers to develop identical tests given the same operational definition (Cronbach, 1987)"(p. 247). Haladyna reiterated the role of clinical practice being critical to competence and were echoed by Tamblyn (1994). Haladyna acknowledged standard setting was based on human judgements. Decisions were often based on an underlying assumption that content experts setting standards can identify the performance of borderline individuals for any item. Haladyna concluded that more conceptual investigations are needed for standard setting procedures.

Literature linking cognitive psychology with setting standards, that is, the thought process used to make standard setting judgements was meagre until Maguire, Skakun, and Harley (1992) suggested that a cognitive model may be more appropriate and coincide with how students solve items. These researchers' contention was that competence was related to clinical reasoning and explained, ". . . procedures for setting standards must be related to the kinds of clinical reasoning processes used to answer the item" (p. 436). Maguire et al. used a continuum model, judgemental approach, to tie cognitive psychology with setting standards. These researchers contended that when competence is viewed as a cognitive concept, a Nedelsky type approach should be used to set the standard because it most parallels students' reasoning. They stated,

> When judges are asked to examine items from the perspective of the candidates and to make their assessments in terms of the quality of reasoning, their judgements appear to reflect the thinking processes of that the candidates use as they attempt to solve the problem. . . . If we accept multiple-choice items as a valid and efficient assessment device for clinical reasoning skills, then the Nedelsky-Gross procedure with its strong validity claims may be the procedure of choice for setting standards. (p. 451)

Triska, Skakun, Maguire, and Harley (1995) investigated setting standards from a cognitive psychology perspective. Five clinicians thought aloud while solving five

multiple-choice items and the results were compared to the protocols of 40 students collected by Skakun (1994). The clinicians then set the standard using the Nedelsky procedure for each item. The two main findings reported were: (1) the clinicians had difficulty in explaining why a distractor may be plausible to a competently reasoning student, and (2) clinicians can take a student's perspective to solve an item (with prompting and reminding). These two studies set the tone for this research project, linking cognitive psychology with psychometrics in medical education.

Cognitive Psychology

Literature pertinent to this study is in two categories, cognitive psychology and cognition in medical education. The concepts and methods to investigate reasoning and solving problems by Chi et al. (1981) were used in medical education problem solving studies (see Meskauskas, 1986). To unpack the ideas this study was based on, the relevant literature in both areas will be discussed, beginning with cognitive psychology. The next subsection deals with how cognitive psychology relates to reasoning processes of clinician judges who are setting standards in medical education.

Using experts and novices for physics problem solving, Chi et al. (1981) found that problem solving difficulties of novices can be attributed mainly to their inadequate knowledge, and not to limitations in their cognitive systems or processing capacities. There were similarities between the architecture of novices' and experts' cognitive systems. Novices showed effective search heuristics when they solved problems using backward-working solutions. Novices were equally competent as experts in identifying the key features in a problem statement. The limitation that novices had was their inability to infer further knowledge from the literal cues in the statement problem. Experts could generate these inferences in the context if relevant knowledge structures were present.

Glaser (1989) said that a beginner's knowledge of a specific domain is erratic,

consisting of isolated bits of relevant information, definitions, and superficial comprehension of central terms and concepts. As individuals learn the basic elements of knowledge become increasingly interconnected and structured within a domain. Knowledge acquisition becomes a sequential transition through a series of iterations. The chunks become chains, and chains are combined with previous knowledge, elaborated upon, then integrated into a higher level. As individuals learn to perform within their range of knowledge, a transitionary zone occurs when the knowledge chunks become interconnected into coherent chains of information. These chains of information are formed into a structured piece of knowledge termed a schemata. Schema are organized hierarchically and suggest more advanced thinking. When individuals become domain-proficient, schemata are retrieved rapidly. Novices work at the surface level of a problem, whereas experts organize their knowledge in terms of schemata that enable them to grasp the structure of a problem and bypass a novice's superficial search pattern (Chi et al., 1981; Reimann & Chi, 1989; Rumelhart, 1980).

Glaser (1989) reported that results across various studies have shown that certain features are typical of proficient performance in experts. Expert knowledge is structured, based on principles, and is well integrated. Experts effectively represent knowledge. Experts use proceduralized knowledge-they know when to use what. Experts have skilled memory, have automaticity, and have developed self-regulatory skills. When experts use knowledge, it is functional and bound to conditions of applicability; experts have integrated basic and advanced components of skill. Their attention alternates between basic skills and higher levels of strategy.

In medical education, Groen and Patel (1989) adopted a schema framework to describe clinical reasoning in medical education. They differentiated between novices (students) and experts (clinician judges) by the strategy each group used to solve a

problem. The strategy novices/students used in problem solving was to identifying facts, definitions, and fragments of information. They tried to account for the facts of a scenario by generating many answers and test them against various hypotheses. This method of reasoning was termed backward-working strategy, and was characteristic of unsophisticated problem solving. As medical students progressed through their curriculum of studies, their knowledge base and clinical performance become integrated and structured. This resulted in elaborate and well-structured schemata being created, the type of schema used in medical diagnosis by clinicians.

Conversely, experts used a forward-working strategy to solve problems. Their schema were well-developed, highly structured, and integrated with practice. The experts' schema enabled them to explain additional facts after completing a main diagnosis (tying up loose ends). Clinicians generated hypotheses quickly, accurately, and offered insights into the clinical presentation by relating the scenario to their medical practice. They knew which information was critical to use, and when to use it. This type of forward reasoning was typical of experts' behaviour when they solved problems within their domain of knowledge.

Lemieux and Bordage (1992) investigated diagnostic thinking using a different approach, structural semantics, and compared the results with the propositional diagnostic work of Patel and Groen (1986). Lemieux and Bordage examined the levels of meaning or semantic axes, both horizontal and vertical (see Lemieux and Bordage, 1992; for a complete explanation). They reported that deep underlying semantic structures determine clinical competence, and concluded by saying, "Well-controlled definitional strategies, that is, the adequate recognition of the substance, form, and effect of the symptoms and signs, are the best means of arriving at a diagnosis." (p. 203)

Schmidt, Norman, and Boshuizan (1990) viewed medical expertise from cognitive

structures they called illness scripts. These scripts "contain relatively little knowledge about pathophysiological causes of symptoms and complaints but a wealth of clinically relevant information about disease, its consequences, and the context under which illness develops." (p. 611). Schmidt et al. described a four-stage theory of clinical reasoning, and provided illness script examples for each stage. When they synthesized the findings, they identified five phenomena dealing with content specificity and content domains, data gathering and expertise, criterion setting, intermediate effects, and expertise in visual domains. They implied that the distinction between experiential and conceptual knowledge in the clinical domain is a fundamental issue and must be acknowledged. They commented on the student's transitionary phases in relation to assessment and the appropriateness of the assessment tool to evaluate the student's skills. The researchers concluded that it is counterproductive to develop rules for diagnosis because much of the diagnosis is a holistic judgement about similar conditions.

Elstein, Shulman, and Sprafka (1990) summarized medical problem solving by presenting a ten-year retrospective, from 1980 to 1990. They suggested that a multipronged, multimethod program of research on clinical reasoning and decision making might shed light on how these processes occur and the best method to use technology to reinforce learning.

Various perspectives of cognitive psychology have been offered to acquire knowledge. Each concept attempts to distinguish experts from novices. Defining the differences in reasoning between experts and novices is difficult because the definitions of expert reasoning are more plentiful in the literature than are definitions of novices. From a cognitive psychology perspective, Chi et al. (1981) said that experts possessed a large body of knowledge and procedural skill. Cramer and Johnston (1991) offered a different perspective by saying that an expert is a person who is experienced at making

predictions in a domain and has some professional or social credentials. Kennedy (1987) argued that the first definition of expertise is derived from the specific task a professional must perform. All these researchers concur that experts appear to have a special ability or skill in a specific domain of knowledge, a broad base of knowledge in a specific domain, and possess several years of working experience in this specific domain.

In medical education, researchers have attempted to identify differences between novice and expert clinical reasoning. Patel and Groen (1989) advocated propositional schemata (1989). Lemeiux and Bordage (1992) thought that structural semantics encompassed a greater depth of knowledge structure for solving problems. Schmidt et al. (1990) offered illness scripts. These researchers approached acquisition of knowledge from diverse perspectives, with the common strand being that expertise may be a set of discrete skills designed to perform specific tasks, rather than generic tasks.

Investigating the concept of linking cognitive psychology with psychometrics in medical education required a framework that encompassed cognition, assessment of cognition, and a standard setting process that would function within these two theories. Before these psychological and cognitive constructs can be addressed, the foundation for investigating the notion of reasoning from a clinician's perspective follows.

Theoretical Models Used in This Study

The theoretical foundation adopted in the present study for investigating the acquisition of knowledge was an information processing theory proposed by Anderson (1983; 1993), the ACT* theory. This was coupled with the cognitive skill assessment model proposed by Glaser et al. (1985) to provide an assessment orientation for determining levels of competent reasoning in medical examinations. Due to the atomistic nature of the item solving behaviours of novices found by Skakun et al. (1994), the

Nedelsky procedure for setting standards was adapted to differentiate competent reasoning from incompetent reasoning when individuals were solving multiple-choice items. Discussion of the ACT* theory, the cognitive skill assessment model proposed by Glaser et al. (1985) and applied by Royer et al., 1993, and Nedelsky procedure for setting standards follows.

The ACT* Theory of Cognitive Skill Acquisition

After reviewing cognitive psychology literature (Chi, et al., 1981), and medical education literature (Elstein et al., 1990; Groen & Patel, 1989; Lemieux & Bordage, 1992; Patel & Groen, 1986; Schmidt et al., 1990), Anderson's (1993) ACT* theory of cognitive skill acquisition was chosen as the framework for this study because it:

1. had been empirically evaluated (Anderson, 1983);
2. pertained to the continuum of the acquisition of medical clinical reasoning skill development (Doolittle & Yekovich, 1994);
3. was used for knowledge-rich domains of learning (Anderson, 1993); and,
4. had been widely recognized as one major force in the theoretical development of the cognitive revolution (Royer et al., 1993).

According to Anderson (1993), cognitive processes used by problem solvers can be viewed as a sequence of internal states of knowledge successively being transformed from declarative knowledge processing and deeper problem representation through to practice and expert problem solving skills. The declarative knowledge becomes very large and well organized while the procedural knowledge becomes more specialized and efficient.

Anderson (1983, 1993, 1994) explained that declarative knowledge, which encodes facts, and procedural knowledge, which encodes the cognitive skill needed to solve problems, are distinct. The initial stages of cognitive skill development are termed the interpretive stages. Declarative knowledge is conceptual and factual and is the result of

acquiring information about something. Novices use declarative knowledge with general procedural knowledge to solve domain-related problems. Eventually, the declarative knowledge becomes elaborate and well organized. As the complexity of problems increases, however, general knowledge is insufficient to reach a solution. When individuals reach a state where there are no adequate solutions, they will search for an example of a similar problem-solving situation and try to solve the problem by analogy to that example. Specific scenarios are recalled and interpreted. Knowledge is retrieved from declarative memory. Individuals rehearse pivotal issues from the situation and substantial verbalization can occur. As the verbalization lessens, the cognitive skills are encoded procedurally. This transition is termed knowledge compilation and is the associative link between the interpretive stage and autonomous stage of procedural cognitive skill acquisition.

Procedural knowledge is encoded in the form of production rules termed condition-action pairs in a series of IF - THEN statements. For example the following item was solved using condition-action pairs by one clinician (FAM04) in the study.

> A 55 year old man complains of severe pain in his left leg. The leg is cool, pale, and pulseless. This situation occurs one week after an anterior myocardial infarction.

From Clinician FAM04's perspective, the condition-action pairs of IF - THEN statements would include:

> IF you have a man with an arterial infarction,
>
> THEN it makes him the highest group risk for thrombosis in the left ventricle and the highest risk for throwing an embolus.

Once knowledge is in this form, it is applied quickly and with great accuracy.

According to Anderson (1993), ". . . a critical factor that determines both the

accessibility of declarative knowledge and the performance of procedural knowledge is the strength of encoding of this knowledge, which basically reflects amount of practice" (p. 41). This strength of encoding grows as practice increases. The factor that controls the strength of learning is knowledge compilation, which is the bridge between declarative knowledge and procedural knowledge. As more practice occurs, procedural knowledge is strengthened and problem solving skills become more efficient and automatic.

If Anderson's model is followed, medical education can be construed in the following way. Medical students' accumulation of knowledge begins with gathering facts, data, and the identifying key features of a specific scenario. They begin this process in the declarative knowledge stages of their medical training after completing two years of basic sciences' courses. The transition between declarative knowledge and procedural knowledge takes place in their preclinical years of medical school where specific medical domain knowledge is acquired by courses, laboratory exercises, and clinical rotations. This transition is further enhanced in the students' clinical clerkships during the last two years of medical school where they practice their clinical skills, strengthening their knowledge compilation and procedural skills. As students' apply their knowledge and procedural skills through practice, they become more efficient. When medical students become practicing clinicians, their domain knowledge is so well integrated into their procedural knowledge that their actions become automatic and refined. Their problem solving skills become more developed, and they begin monitoring their own behaviour and decisions, without external feedback. Clinicians appear confident in their problem solving skills and their diagnoses are usually appropriate and accurate.

Development of cognitive ability with the ACT* theory of acquisition of medical knowledge and skills can be viewed from a continuum perspective (Figure 2.01). When individuals begin their medical training, two years of study are devoted to basic sciences

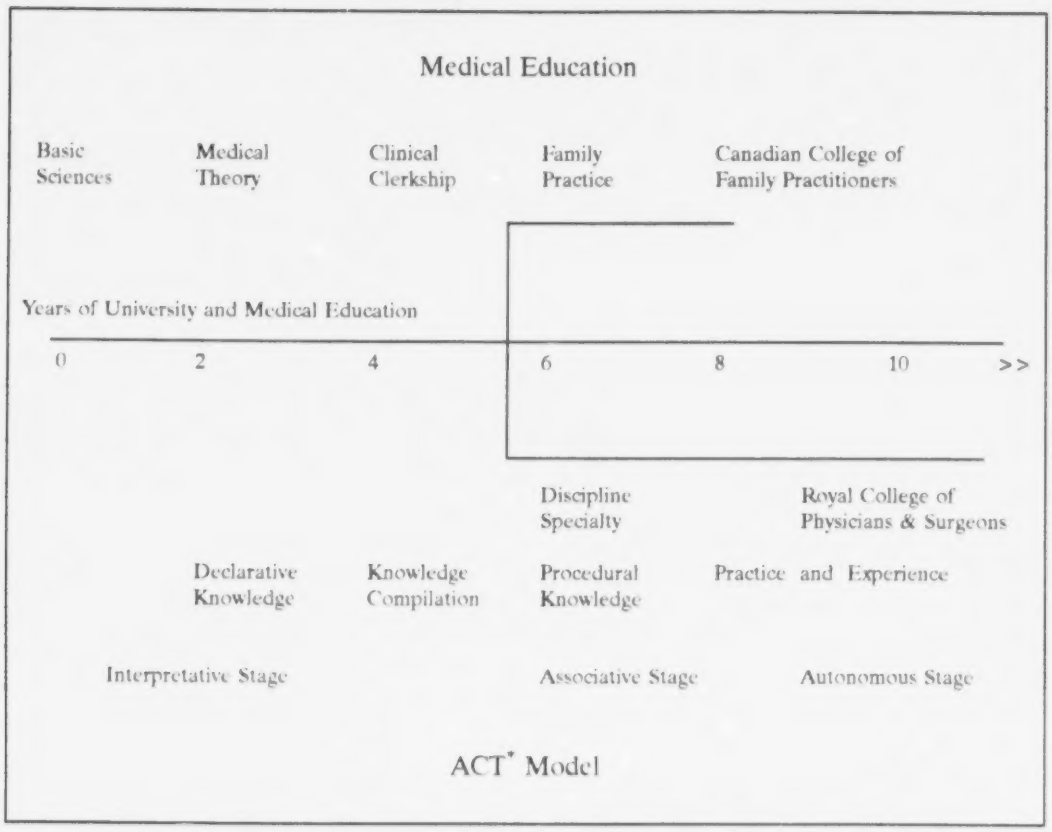


Figure 2.01. Continuum of Medical Education and the ACT* Model.

before admission to medical school. After students are admitted to the Faculty of Medicine, their preclinical two years of instruction take place in a classroom and laboratory setting. They are given facts, case studies, and basic information needed to function in the clinical arena. During the next two years of clinical clerkships, students begin integrating their facts, data, and examples of clinical situations with observing and practicing basic skills of medicine. After their fourth year of medical school, students continue to prepare themselves to attend to patients' needs and diagnose illnesses by entering residencies in their area of interest. When this period of transitions is completed, these physicians establish their own medical practices' and continue to apply

their procedural skills and problem solving abilities on a daily basis.

As physicians continue to use their skills, their behaviours become well developed, efficient, and automatic. They begin to view clinical scenarios from concepts and underlying principles, in contrast to students who view the scenarios from a data and fact basis. When clinicians are unfamiliar with a domain of knowledge, they revert to the lower levels of cognitive development and reason, similar to students. The depth of domain specific knowledge varies from individual to individual, as does problem solving skill.

The Cognitive Skill Assessment Model

The cognitive skill assessment model proposed by Glaser et al. (1985) functions in concert with the ACT* theory and provides an assessment framework to determine levels of competent reasoning in medical situations. Assessment of competency requires clinician judges to project how far along the continuum of medical education the student should be. This task requires clinician judges to assess the level of clinical reasoning skills of the medical students about to enter residency. According to Glaser et al., these skills tend to be sequential in a domain-specific area of knowledge. These skills span several dimensions, from knowledge differences of novices to experts, with intermediate stages, and transitions from level to level. The intermediate stages and the nature of the transformation of skill from one phase to another are difficult to pinpoint. An explanation of the assessment of knowledge along the dimensions from novice to expert follows.

Beginners' knowledge is fragmented. It consists of isolated facts, definitions, and a superficial understanding of the domain vocabulary. The degree of fragmentation and structuredness and accessibility of the interrelated chunks of knowledge suggests six levels of assessment. At the first level of assessment, knowledge organization and structure can

be examined to determine the elements and components in a domain of knowledge. The objective is to determine the interconnections between the chunks of information and fragments of knowledge.

Second, one can assess the depth of problem representation an individual possesses. Novices typically recognize the surface features of a problem or task. More proficient individuals identify inferences, principles, and concepts imbedded below the surface structure. Individuals who immediately recognize underlying principles solve problems rapidly by spending little time on details. They arrive at the correct answer quickly.

Third, the quality of mental model can be investigated. Individuals develop mental models of scenarios and situations consistent with their domains of knowledge. The nature of these representations is determined by the tasks necessary to execute the performance of skills at a specific level. As tasks become more complex, the model is amended to incorporate these new skills. The mental model not only indicates the level of task complexity but also the level of cognitive ability needed to solve the problem.

Fourth, the efficiency of carrying out procedures can be viewed. As individuals' mental models become more refined, tasks are performed more effectively. Well-practiced procedures are important for understanding and comprehension of a scenario. At this level, assessment should focus of the relationship between understanding the requirements of a task and performing the task.

Fifth, automaticity should be present. At higher levels of cognitive development, previously learned basic skills are practiced enough so that they become automatized and are performed with little conscious attention. When performing complex tasks, individuals' attentional demands are taxed simultaneously. As this occurs, the efficiency of the overall task is compromised. The criterion for assessment is whether the

automaticity of the basic process has progressed to a point where the subtasks have minimal interference on the total performance.

Sixth, metacognitive activities should be present. At this level, individuals monitor their performance and identify strategies for solving problems using procedural skills. Individuals reflect on and control their performance. They know what they know or don't know. They can predict the outcome of their performance. They can plan their activities in advance and efficiently apportion their time. They can check and monitor the outcomes of their problem solutions and attempts to learn. These regulatory skills develop with experience and practice. Evaluating metacognitive skills could be an important facet of predicting problem-solving abilities that result in learning.

Glaser et al. (1985) explained that these skills tend to be indexed in a domain of knowledge. Because these skills are not discreet but on a continuum, identifying assessment tools appropriate for each level is difficult, particularly at the transitions points and at higher levels of cognitive development.

Differentiating among the higher levels of knowledge is demanding because experts' cognitive structures are well developed. The speed of processing knowledge is so fast that the borders among levels are indistinct, particularly between efficiency of procedures and automaticity. Metacognitive activity can be difficult to identify when individuals change their train of thought, for example when they generate a hypothesis, then consider peripheral issues and offer another solution to a problem. In responding to a multiple-choice item, experts might read an item stem and generate a hypothesis. If after thinking further about the scenario, they change the hypothesis, did the experts just change their minds, or did they reconsider the information presented in the stem? After rethinking the scenario, perhaps they did incorrectly view the situation. Are they monitoring their behaviour based on experience, or did they consider additional

information retrieved from their elaborated schemata and just change their opinion? This situation is difficult to classify in the assessment of skills hierarchy.

Royer et al. (1993) offered a strategy to assess cognitive development using the theory that Glaser et al. (1985) proposed and linked it to Anderson's (1983) ACT* model for acquisition of knowledge. These researchers suggested that knowledge acquisition runs parallel to the assessment of clinical reasoning skills along a continuum. For the purposes of this study, the researcher proposed that knowledge acquisition, assessment of these skills, and medical education could be overlaid on the same continuum as shown in Figure 2.02.

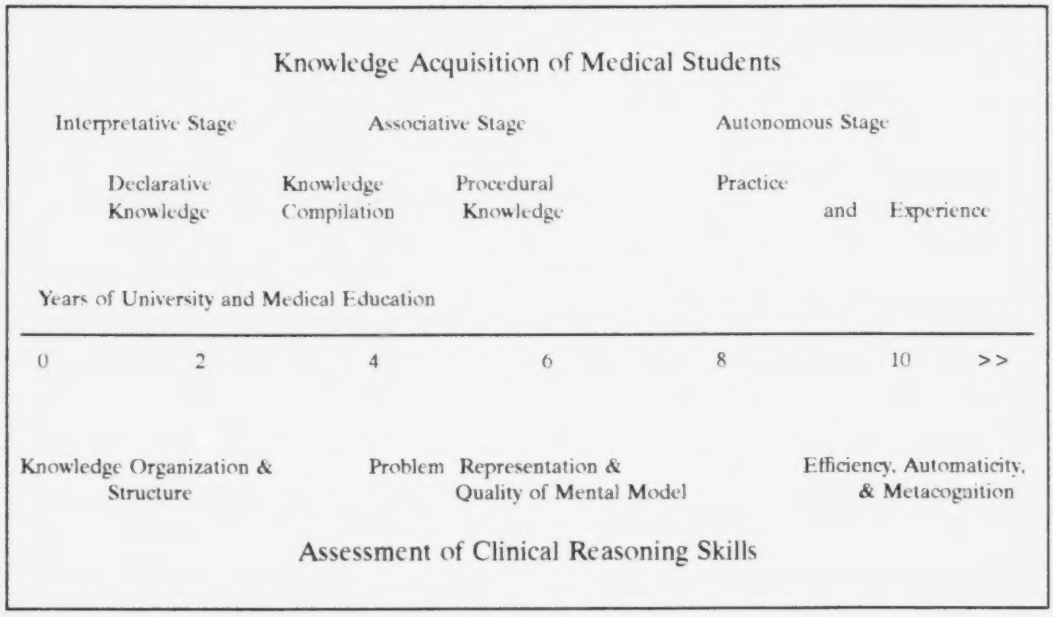


Figure 2.02. The ACT* Model, Assessment of Reasoning, and Medical Education.

The starting point is admission to university, but there is no closure because individuals continue learning over their entire professional lives. Conceivably,

professionals' competence may be assessed regularly by their licensing agencies until they choose not to practice their profession.

The Nedelsky Procedure for Standard Setting

Nedelsky (1954) designed a standard setting procedure specifically for multiple-choice items. He explained,

> The passing score is to be based on the instructor's judgement of what constitutes an adequate achievement of the part of a student and not on the performance by the student relative to his class or to any other particular group of students. . . . The standard is absolute if it can be stated in terms of the *knowledge and skills* a student must possess in order to pass the test. (p. 3)

Nedelsky clearly stated that a failing or barely passing student was a statistical abstraction, but offered no advice on how to identify this individual. He went on to explain the technical aspects of setting the passing score using a four-point system within an item with five alternatives. When discussing the refinements of the technique, he cautioned the reader about three issues that affect validity: time to make judgements, knowledgeable judges, and training the judges in objective item testing. These were his suggestions, but were not expanded in this article. In the results, Nedelsky reported that the instructors in the study felt that the item difficulty depended not only on selecting the keyed response, but also on the quality of the distractors.

Nedelsky's focus was on the technical aspects of standard setting, that is, setting an absolute passing score on a series of items in an objective test. He used the notion of a barely passing student, acknowledged the validity of the decisions, but ignored the theoretical structure of both concepts.

Summary

The thrust of this study was to examine the validity of clinicians' judgements in setting standards by using a distractor approach when placed in a cognitive psychology framework. Going one step further, setting standards for reasoning must include the assessment of reasoning at the distractor level. Nedelsky did not explicitly state this, however, this study was based on the Nedelsky procedure or Nedelsky's theory. Nevertheless, for convenience, the procedure for setting standards was labelled the Nedelsky type procedure (NT) and the theory surrounding it, the Nedelsky type theory. Each distractor is addressed individually by each judge (atomistic approach), without group discussion.

Standards should be set in relation to a position appropriate to where a student's standard of performance should be on the knowledge continuum (Anderson, 1983) and the assessment continuum (Royer et al., 1993). The theory of expertise noted earlier suggests that there are qualitative differences in reasoning between experts and novices. This implies that clinician judges and students will think differently about why a distractor is wrong or why the keyed response is right. Clinicians should have a more elaborate and well-structured knowledge domain as compared to students. Student would rely on facts, discrete bits of information, and definitions to reason to the keyed response. Because of these different knowledge structures, both groups should approach item solving using different strategies.

If the standards are to be valid, they should be set appropriately for the student's level of reasoning on the knowledge continuum. According to the ACT* theory, judges need to be able to move down the continuum to view distractors with all of the complexities as a student would, and make judgements about what a student ought to do to be deemed competent at that level. A critical issue then is-can the clinician judge

perform this task?

In the present study, clinician judges carried out the Nedelsky procedure for each item. The specific instructions are shown in Appendix B. Rather than discriminating between nonfailing and failing students, clinicians were instructed to examine each distractor, then decide if a competently thinking Phase III student, who has just completed his or her medicine rotation, should eliminate it as incorrect. To unpack the judges' psychological constructs on which they made their decisions, judges thought aloud as they solved the items and set the standard. Nedelsky's procedure for calculating passing scores is external to this study and, therefore, was not addressed.

The next chapter contains a description of the populations in the study, the selection and description of the multiple-choice items, outlines the data collection procedures, and presents the results of a preliminary analysis of competence and expertise. The chapter concludes with the implication of the analyses on the remainder of the analyses in this inquiry.

Chapter 3 - Method and Preliminary Analysis

Introduction

The first section in Chapter 3 discusses the relevance of a preliminary analysis, then describes the selection of the clinicians, the characteristics of the students, the selection and description of the multiple choice items in the study, and outlines the data collection procedures. The second section presents the results of the preliminary analysis. Chapter 3 ends with a summary of preliminary analysis and the implications of the results on the clinicians' judgement comparisons.

Relevance of Preliminary Analysis

A preliminary analysis was needed to set the framework for the primary analysis, which was to explore the validity of judgements in standard setting using a distractor based approach. The preliminary analysis was conducted to determine the competency of the clinicians in this study, and to assess the degree of expertise shown by competently reasoning clinicians and students.

A premise of this study was that to set credible standards of student performance on a particular item, clinician judges must possess a level of cognitive competence that is commensurate with the declarative and procedural knowledge implied by that item. For this study, competence with respect to an item was defined as choosing the keyed option.

Further, according to the theory of standard setting used in this study, competent clinicians should be able to examine the item from the perspective of a student. In the standard setting task, judges look at each alternative in turn and indicate whether or not students should know the alternative to be incorrect. According to Anderson's (1983, 1993) ACT* theory , judges are generally further along the continuum of expertise than

students. The distance between the clinicians and students may influence the judgements the clinicians make. Simply put, expert judges may not be able to view the item from the perspective of a competent student and as a result the standards they set may be unrealistic. To see if this could be a problem, clinicians and students were classified according to their level of expertise using six dimensions along which expert and novice reasoning can be assessed (Royer et al., 1993). It was thought that if differences were present in expertise between the clinician and student groups, subsequent analyses would be needed to explore the influence of expertise on judgements made during standard setting.

Selection of the Clinicians

The criteria chosen for the present study are consistent with those used by the Medical Council of Canada's examination committee. It was felt that using a sample that was typical of standard setting panels for medical licensure examinations would enhance the validity of the study. Physicians chosen for this study met the following the criteria. Each clinician was:

1. provincially licensed to practice medicine in Alberta,
2. a physician with more than five years of clinical practice experience,
3. a physician with a minimum of five years of current experience in medical education, thereby having a working knowledge of the medical education curriculum, and
4. a physician who had contact with medical students in his/her daily practice.

An attempt was made to recruit physicians from the Faculty of Medicine at the University of Alberta. Of the 16 physicians approached, 12 agreed to participate in this study. There were four from family practice, two each from general internal medicine,

infectious diseases, and pulmonary medicine, and one each from haematology and gastroenterology.

Characteristics of the Students

Rationale for Using the Students Protocols From Skakun's Study

Student protocols, obtained from Skakun (1994), were an essential component of this study and were used in two ways. First, the students' protocols were used to compare the level of expertise in competent clinician and competent student reasoning when they solved the same multiple-choice items. Second, the students' protocols were compared with the clinicians' perceptions of how students reasoned to solve seven items.

Characteristics of the Students

In previous studies, Skakun (1994) and Skakun et al. (1994) analyzed think aloud protocols from 40 third and fourth year medical students as they solved 30 internal medicine multiple-choice items. Students had completed a clinical rotation at the University of Alberta Hospitals or another affiliated hospital before participating in Skakun's study. As previously explained, only competent student protocols were used in the present study. Details of the selection process for the students are found in Skakun (1994).

Description of the Multiple-choice items

Ten multiple-choice items (Appendix A), eight positively worded and two negatively worded, were selected from the University of Alberta, Department of Medicine item bank. Eight of these items were chosen by Skakun (1994) and used in the present study to compare the clinicians' perceptions of student reasoning with the students' protocols. Also, these items represented various disciplines in medicine and diverse levels of item

difficulty.

From the time the items were administered to the students in Skakun's (1994) study to the time of the present study, some items in the bank were revised. Item 773 was revised after it was administered to the students This revision resulted in a major difference in reasoning between the clinicians and students and the revised version was inadvertently used in the present study. Because of this, only the clinicians' protocols for item 773 were used.

Under Dr. C. Harley's (Associate Dean of Undergraduate Medical Students) guidance, two items were chosen from the Department of Medicine at the University of Alberta to represent two other disciplines in internal medicine, haematology and rheumatology. In total, 10 items were used for the present study, eight from Skakun's (1994) study, and two from the item bank.

Items used in this study had been previously classified by the clinicians in the Department of Medicine at the University of Alberta, Faculty of Medicine. According to that categorization scheme, recall items require students to remember facts and details learned in a preclinical or clinical situation (item 060). Comprehension items require students to understand the content of the stem, and to use this information to select the keyed alternative (item 347). In diagnosis items, a clinical situation is described, and students are required to formulate a diagnosis from the information in the stem (items' 031, 332, 582, & 733). For test interpretation, various test results are given in the stem of the item, and students are required to interpret the results (item 267). Management items require students to construct a problem space for a diagnosis for the condition described in the scenario, then decide how to manage the symptoms (items 317, 771, & 773).

Included in the study were two negatively stated items (items' 060 & 267). In the

negatively stated items students are required to select the alternative that doesn't belong and the word **EXCEPT** was highlighted to draw their attention to the phrasing in the stem. They were used in this study because these items are routinely used on examinations and clinicians set standards on negatively stated items.

Data

All of the clinicians who consented to be in the study were sent preliminary materials to review. The information package included an overview of the interview, a consent form, and an example of an item being scored using a think aloud strategy (Appendix B).

Each clinician judge was interviewed independently without interruption. Every interview was recorded and completed in one session. The total time required to gather data from each clinician judge was between two, and three and one-half hours.

At the beginning of every interview the procedure was explained in detail, and clinicians were asked to sign the consent form to participate in the study. To ensure that the clinicians understood the process, each clinician solved a practice item. As every clinician progressed through the stages of the procedure for the practice item, additional information and guidance were given where needed, and questions about the procedure were answered.

Each clinician's think aloud interview consisted of five stages: (1) solving the multiple-choice item, (2) setting the standard using the NT procedure, (3) explaining the misconceptions for each distractor in each item, (4) rating the competency of a student if a distractor was chosen as a correct response, and (5) explaining how a Phase III student would reason to select the keyed response. Details of each stage (except stage 3) will be provided in the analysis section. All stages were completed for one item before proceeding to the next one.

The items were presented in the same order to all clinicians (Appendix A). The negatively stated items were placed together in the third and fourth position for the interview. This was done intentionally because it was thought that if these two items were placed at the end of the interview, clinicians' judgements might be inconsistent because of fatigue.

All interviews were recorded on audiotapes and field notes were taken. Transcriptions were done immediately after the interviews. After each interview was transcribed, it was verified by listening to the audiotape while reading the protocol for accuracy.

Preliminary Analysis

Competency

As noted earlier, competent reasoning with respect to a particular item was operationally defined as selecting the keyed option. In Stage 1 each clinician was presented with an item and asked to choose the correct alternative, thinking aloud as he/she solved the item.

For convenience, the clinicians who chose the keyed alternative will be referred to as Group 1 (G1) and those who did not will be referred to as Group 2 (G2). Membership in G1 and G2 varied for each item. The number of clinicians in G1 varied from 12 to five across the 10 items. One clinician chose to omit item 267 because the discipline the item represented was outside the area of his/her knowledge. One clinician did not choose the keyed response for item 773, and subsequently treated the option as the keyed response. Therefore, this clinician was not used for the distractor analysis in Chapter 4. Table 3.01 shows clinicians' responses relative to the keyed alternative for each item.

Table 3.01

Summary of the Clinicians' Who Chose the Keyed Response

Item Number	Group 1	Group 2
771	12	0
031	11	1
317	11	1
347	11	1
582	11	1
773	9	3
060	9	3
733	8	4
267*	5	6
332	5	7

Note: * One clinician omitted this item because of the content, therefore, n = 11.

Assessment of Reasoning

Once competence was established, identification was removed from all of the competently reasoning clinician and student protocols. The clinician and student protocols were randomly interspersed for each item and the items were sorted by the researcher into two categories based on the reasoning displayed in the protocol. The clinician and student protocols were classified as expert and novice according to six dimensions shown in Tables 3.02 and 3.03.

Verification. The results of the dichotomous sort for the ten items were validated by having an independent sort done by another rater for all of the protocols. If agreement was not reached on a particular protocol, both raters discussed the reasoning process using Tables 3.02 and 3.03 in an attempt to reach agreement. After the first sort by both raters, the interrater agreement was 85 percent. Both raters discussed the remaining protocols, and agreement was reached on all protocols for all items. The raters agreed that one protocol in item 771 posed a problem because the level of cuing provided by the

Table 3.02
Description of Cognitive Assessment of Experts*

Dimension	Description
Knowledge organization and structure	Individuals activate large chunks of knowledge within a skill domain.
Depth of problem representation	Individuals perceive problem from abstract principles that subsume the surface and related problems.
Quality of the mental model	Individuals have complex and elaborate model developed consistent with the problem/situation. They have the presence and sophistication to imagine/envision the problem/situation within a domain that guides their performance.
Efficiency of procedures	Individuals follow a solution path that eliminates any unnecessary steps in a fail-safe manner; less time is taken to solve the problem with fewer steps.
Automaticity of performance	Individuals can handle many aspects of performance in an automatic and nearly load-free manner. Leaves a certain amount of cognitive capacity available for performing other activities.
Metacognition	Individuals have self-regulatory skills and self-management skills which they use to reflect upon their performance by: knowing when or what they know or do not know; predict the correctness or outcome of their performance; plan ahead and efficiently apportion their time; and check and monitoring the outcomes of their solution.

Note: *Glaser et al. (1985); Royer et al. (1993).

Table 3.03

Description of Cognitive Assessment of Novices*

Dimension	Description
Knowledge organization and structure	Individuals activate isolated facts, definitions, and concepts.
Depth of problem representation	Individuals perceive problem in terms of the particular elements present in the problem. They recognize surface features of a problem or task situation.
Quality of the mental model	Individuals have difficulty in imagining/envisioning the problem/situation within a domain.
Efficiency of procedures	Individuals reach a correct solution by systematically following a fail-safe sequence of steps.
Automaticity of performance	Every aspect of an individual's performance is frequently performance based on conscious reasoning processes. Their ability to process information other than that associated with the immediate problem is nil.
Metacognition	Individuals are less proficient in monitoring their performance and less successful in applying the skill they do possess.

Note: *Glaser et al. (1985; Royer et al. (1993).

interviewer was excessive for one clinician. After rereading and discussing the protocol, a decision was made to remove this protocol from the sort.

Results. Competent G1 expert clinicians were labelled G1E and the G1 novice clinicians were labelled G1N. Competent expert students were labelled ES and novice students were labelled NS.

As explained earlier, it was thought that if differences were present in expertise between the clinician and student groups, subsequent analyses would be needed to explore the influence of expertise on judgements made during standard setting. Table 3.04 shows that expert and novice reasoners were found amongst both the competent clinicians and competent students.

Table 3.04

Results of the Cognitive Assessment

	Clinicians (G1)				Students			
Item #	n*	G1E		G1N	n*	ES		NS
771	11	10	(90.9)**	1	22	5	(22.7)**	17
317	11	9	(81.8)	2	32	19	(59.4)	13
347	11	7	(63.7)	4	21	16	(76.2)	5
582	11	11	(100.0)	0	30	22	(73.3)	8
733	8	3	(37.5)	5	18	7	(38.9)	10
267	5	3	(60.0)	2	3	2	(66.7)	1
332	5	4	(80.0)	1	22	14	(63.6)	8

Note: * Number of individuals who chose the keyed response;
** Proportion of total n's.

The ratio of expert clinicians (G1E) to novice clinicians (G1N) varied across items, as did the ratio of expert students (ES) and novice students (NS). Two groups of

particular interest were the expert clinicians and expert students because, theoretically, the clinicians should be further along the continuum of expertise than students, so there should be proportionally more expert clinicians.

There were proportionally more expert clinicians than expert students for four items from the disciplines of gastroenterology (two), cardiology (one), and pulmonary medicine (one). These items were classified as management (two) and diagnosis (two) items. The results showed that there were only minimal differences in favour of the expert students for three items. These three items were from the disciplines of neurology, pulmonary medicine, and endocrinology, and classified as diagnosis, comprehension, and test interpretation, respectively. The differences were significant for only two of the items, one from gastroenterology (management) and from cardiology (diagnosis).

To decide if there were significantly more expert clinicians than expert students, a Z-test (one-tail test at $\alpha = 0.05$) for the difference between the proportions of clinicians and students was done. The results (see Table 3.05) suggest that if experts do have difficulty

Table 3.05

Z-test to Determine the Difference Between the G1E Clinicians and Expert Students

Item #	Discipline	Classification	Difference Between G1Es and ESs
771	Gastroenterology	Management	3.70*
582	Cardiovascular	Diagnosis	1.93*
317	Gastroenterology	Management	1.43
332	Pulmonary Medicine	Diagnosis	0.60
267	Endocrinology	Test Interpretation	-0.17
347	Pulmonary Medicine	Comprehension	-0.19
733	Neurology	Diagnosis	-0.34

Note: * One-tail test significant at $\alpha = 0.05$.

judging the item from a student's perspective, this may be most problematic for items 771 and 582, and possibly for 317 and 332.

It was thought that clinician expertise may influence standard setting judgements. To examine the extent of the possible influence for each item, the binomial test was used to test the null hypothesis that the frequency of clinician categories as experts and novices was the same. This hypothesis was rejected (one-tail at $\alpha = 0.05$) for items 771, 582, 031, 347, and 317 showing that there were significantly more expert clinicians than novice clinicians for these items (Table 3.06). The difference between the G1E's and G1N's was

Table 3.06

Frequency of G1 Clinicians' Judgements

Item Number	G1E	G1N	Total in G1
771*	10‡	1	11
582*	11	0	11
031*	10	1	11
347*	9	2	11
317*	9	2	11
773	6	3	9
060	5	4	9
733	3	5	8
332	3	2	5
267	3	2	5

Note: * One-tail test significant at $\alpha = 0.05$;
‡ For subsequent analyses G1E = 10

not significant for each of the remaining items. The results suggest that if expertise influences the way in which clinicians set standards, it might be most apparent in items 773, 060, 733, 332, and 267 where the proportions in the two groups are similar.

Summary

The objective of the preliminary analysis was to establish a framework for exploring the validity of judgements in standard setting using a Nedelsky type approach. The results show that one cannot assume that the criteria for selecting clinician judges will necessarily yield clinicians who are competent with respect to all items.

Once competence was established, the clinician and student protocols were classified as expert and novice according to the dimensions outlined by Royer et al. (1993). The purpose for classifying competent reasoners, both clinicians and students, was to determine if clinicians were further along the continuum of expertise than students. Expert clinician reasoners and expert student reasoners were identified for all items. Novice clinician reasoners were found in all but one item.

When the proportion of expert clinicians was compared to expert students, there were significantly more clinicians than students for only two items. One of the items was classified as management in the discipline of gastroenterology (item 771). Content analysis of the protocols from this item revealed that expert clinicians relied heavily on their procedural knowledge and clinical experience to solve this item. Expert clinicians used similar strategies to solve item 582. In the items where differences between expert clinicians and expert students were smallest (items 267, 347, & 733), the items were factual and required more of a declarative knowledge base. Content analyses of the protocols suggested that differences in expert clinician and expert student reasoning may be attributed to the quality of the clinical reasoning, the discipline, and item classification.

Using the binomial test, the frequencies of expert clinicians and novice clinicians were compared for each item. The results showed that there were significantly more expert clinicians than novice clinicians in five of the 10 items. Content analyses of the protocols revealed that the calibre of clinician reasoning depended on the content of the

item, and the frequency of the clinical scenario arising in daily practice.

Implications of the Results on the Judgement Comparisons

To set valid standards, clinicians must be able to identify indicators of competent reasoning. If they select the keyed option when responding to an item, they may not be able to identify options a competently reasoning student might eliminate as incorrect, and to explain how students reason to select the keyed response. The positions that clinicians occupy in the knowledge continuum in relation to competently reasoning Phase III students may influence the appropriateness of the standard that they set for a particular item. These issues form the foundation of the interpretation of the analysis of decisions in setting standards (Chapter 4), and the analysis of the clinicians' perceptions of how competently reasoning students solve items (Chapter 5).

Chapter 4 - Judgement Comparisons

Introduction

Chapter 4 is divided into five sections. The first section presents the rationale for comparisons of clinicians' judgements, and for using an alternative distractor type standard setting process. In the second section, the analyses and results of the NT procedure for elimination of distractors are presented for G1 and G2. In the third section, comparisons of judgements made within G1 (G1E vs. G1N) are described and the implications for the remainder of the study are presented. In the fourth section, the analyses and the results of the comparison of judgements between the NT procedure judgements and the alternative distractor approach within the G1 clinicians are described. Chapter 4 ends with a summary of the results and the implications for a theory of standard setting.

Rationale for Comparisons of Clinicians' Judgements

The NT procedure for setting performance standards in multiple-choice items is essentially a judgemental process in which clinicians identify indicators of competent and incompetent reasoning. More specifically, the procedure calls for identifying distractors that students at a particular level should know are wrong. As shown in Chapter 3, not all clinicians in this study solved each item correctly. For each item clinicians were labelled G2 if they did not solve the item correctly. Of the clinicians who selected the keyed option (G1 clinicians), there were two subgroups, expert reasoners (G1E) and novice reasoners (G1N) for nine of the 10 items.

Two issues required resolution before the clinicians' judgement thought processes could be examined. First, are judgement differences present between G1 and G2

clinicians when they eliminate distractors using the NT procedure? To explore this question, comparisons of NT passing scores and the patterns used to eliminate distractors were made. It was assumed that if only small differences were present in both passing scores and patterns of elimination of distractors between the two groups, then G1 and G2 clinicians could be combined for subsequent analyses. If differences were present, then it would cast doubt on the use of clinician judges who do not choose the keyed response.

Second, are judgement differences present between G1E and G1N clinicians when they eliminate distractors using the NT procedure? To look at this issue, comparisons were made between the two groups using the passing score and the patterns of elimination. If differences between the groups were not great, then they could be combined for the qualitative analysis of standard setting behaviour.

Rationale for Using an Alternative Distractor Based Standard Setting Approach

An indicator of convergent validity is the consistency of judgements produced by the NT procedure and an alternate distractor based approach. The alternate rating approach, developed for the present study, was based on a competency rating scale (CRS). It was designed to elicit clinician judgements of the level of competence exhibited by students who chose the particular distractor. This competency rating approach is explained in the third section of this chapter. If the clinicians' judgements were consistent between the two methods, it would be convergent evidence in support of the validity of the NT procedure.

The Clinicians' Judgements Using an NT Distractor Based Procedure

A clinician's judgement on each distractor for a given item was the unit of analysis for the G1 versus G2 comparisons. The judgement in each clinician's protocol for each

distractor was entered on a spread sheet, tallied, and checked. The data consisted of ones and zeros indicating whether the clinician thought that a competently reasoning student should reject the alternative as being incorrect (1) or might reasonably entertain the alternative as being correct (0). The keyed alternative is not part of the NT procedure, and was not judged.

For each subgroup, NT passing scores were calculated for each item. The consistency of judgements within subgroups of clinicians were summarized by determining the average number of agreements between pairs of clinicians.

To calculate the NT item passing score, the following formula was used:

$$\text{Passing Score for Item } i = \left(5 - \sum_{j} p_{ij} \right)^{-1}$$

where p_{ij} is the proportion of judges in a subgroup who eliminated distractor j for item i. The number of alternatives was five (four distractors & one keyed option) for all items.

To assess the degree of agreement, the following procedure was used. For each item, a clinician's judgement was expressed as a vector of values. The values in the vector were either 1 or 0 depending on whether the clinician thought a competently reasoning student should eliminate a distractor as incorrect (1) or the distractor should not be eliminated (0). The keyed alternative (-) was not used in the agreement analysis. For example the vector (1, -, 1, 0, 1) would indicate that the clinician thought that competent students should eliminate options a, c, and e, but not option d. Option b is the keyed option. The vectors of pairs of clinicians were compared and the number of agreements was calculated. For example, vectors (1, -, 1, 0, 1) and (1, -, 1, 0, 1) show the maximum possible agreement of four, whereas vectors (1, -, 1, 0, 1) and (1, -, 0, 0, 0) show partial

agreement of two. The possible values of the agreement score are zero, one, two, three, or four.

To calculate the average number of agreements within a subgroup, the agreement values were calculated across all possible pairs of clinicians and averaged. An average value of 3.0 or greater was considered excellent within group consistency. Consistency across two groups was calculated by taking a clinician from each group and calculating the agreement score. This was repeated for all possible combinations of pairs of clinicians (with one from each group) and the agreement scores were averaged.

The number of clinicians varied in each subgroup, with six items having zero or one clinician in G2, and three items having zero or one clinician classified as novice reasoners (G1N). Because agreements cannot be calculated when there are fewer than two clinicians in the group, these items have no values for the average within subgroup agreement.

Comparison of the NT Judgements Between G1 and G2

The passing scores set by clinicians in G1 and G2 are shown in Table 4.01. The number of clinicians in G2 is two or fewer for five of the items, and so it is difficult to make definitive claims about the difference in passing scores set by the two groups. However, in eight of the nine items where some comparison was possible, the G2 clinicians set a lower passing score than the G1 clinicians. The mean passing score for the nine items was 0.478 for G1 and 0.370 for G2. In summary, clinicians who selected the keyed answer set higher passing scores for the item than those who did not.

The agreement analyses carried out on the NT judgements within G1, within G2, and between G1 and G2 are shown in Table 4.02. Four of the items (031, 060, 267, & 582) average agreements of 3.00 (within rounding) or above for G1. Only two items (060 & 733) reached this level for G2, however for G2 there were five items on which there were

Table 4.01

Passing Scores Set by G1 and G2

Item	G1		G2	
267**	0.625	(5)*	0.278	(6)
060	0.600	(9)	0.500	(3)
582	0.524	(11)	0.500	(1)
317	0.524	(11)	0.500	(1)
771**	0.524	(11)	-----	(0)
733	0.500	(8)	0.400	(4)
031	0.478	(11)	0.333	(1)
773**	0.360	(9)	0.286	(2)
332	0.333	(5)	0.333	(7)
347	0.314	(11)	0.200	(1)
Mean	0.478		0.370	

Note: * Number of clinicians are shown in parentheses;
** Total number of clinicians is 11. See Chapter 3.

zero or one clinicians within the G2 group. The proportion of items that reached the "criterion" was the same in the two groups. There was only one item (060) upon which both groups showed agreements of 3.00 or greater. This suggests an interaction between groups and items. For only one item (582) the between group agreement exceeded 3.0, but for that item, there was only one clinician in the G2 group. The mean agreement across all items for which agreement could be calculated is shown in Table 4.2. When the mean is calculated only on items for which both groups have data, the values were 2.60 and 2.46 for G1 and G2 respectively.

The results of the two analyses show that G1 sets substantially higher standards than G2, and G1 is slightly more consistent than G2. Moreover, the cross group comparison suggests that even in a situation like item 060, where each of the two groups has reasonably high consistency, the between group agreement of 2.56 shows that they differ from one another.

Table 4.02

Agreement of the NT Judgements Within and Between G1 and G2 Clinicians

Item	Average Number of Agreements Between Pairs of Judges Within				Between
	G1		G2		G1 & G2
031	3.13	(11)*	--	(1)	2.73
060	3.00	(9)	4.00	(3)	2.56
267	3.00	(5)	2.00	(6)	1.68
582	2.98	(11)	--	(1)	3.36
771	2.76	(11)	--	(0)	--
317	2.76	(11)	--	(1)	2.82
733	2.64	(8)	3.00	(4)	2.75
347	2.44	(11)	--	(1)	2.18
332	2.40	(5)	2.29	(7)	2.51
773	1.94	(9)	1.00	(2)	2.17
Mean	2.71		2.46		2.53

Note: * Values in parentheses are the number of judges in the group.
-- The number of judges was one or less in the group, therefore comparisons could not be made. These are not included in the mean.

Since the focus of this study was to explore the validity of competently reasoning clinicians' judgements in standard setting, only the G1 clinicians' judgements and perceptions were used for the remainder of the analyses in this study. This decision was supported in two ways. First, the previous analyses showed there was a difference in the passing score set by G1 and G2 and there was greater consistency of judgements within the group that chose the keyed alternative than among those who did not. The implication is that standards set by clinicians who did not choose the keyed response might be less valid than standards set by clinicians who did select the keyed response.

Second, if competent clinicians are to examine an item from the perspective of a competently reasoning student, then only competently reasoning clinicians would know

which alternatives students should know are incorrect. Clinicians who did not select the keyed response were not considered competent reasoners for a given item.

Comparison of the NT Judgements Between G1E and G1N Clinicians

The passing scores set for the 10 items by G1E and G1N are shown in Table 4.03.

Table 4.03

Passing Scores Set by G1E and G1N

Item	G1E		G1N	
031	0.476	(10)*	0.500	(1)
060	0.556	(5)	0.667	(4)
267	0.600	(3)	0.667	(2)
582	0.524	(11)	--	(0)
771	0.524	(10)	0.500	(1)
317	0.526	(9)	0.400	(2)
733	0.563	(3)	0.417	(5)
347	0.310	(9)	0.333	(2)
332	0.429	(3)	0.250	(2)
773	0.375	(6)	0.333	(3)
Mean	0.488		0.452	

Note: * Numbers of clinicians are shown in parentheses.

For item 582, all of the clinicians were classified in the expert group. Across the remaining 9 items, the mean passing score set by G1E was higher than that for G1N (0.488 vs. 0.452). For five of the items, the passing score set by the G1N's was higher than that set by the G1E's. The differences in passing scores set by the G1E's and the G1N's is not as clear as the differences found for G1 versus G2.

Table 4.04 shows the average number of agreements between pairs of clinicians within and between groups. The means of the average within group agreement were 2.78

Table 4.04

Summary of the NT Judgements Between G1E and G1N Clinicians

Item	Average Number of Agreements Between Pairs of Judges Within G1E	Within G1N	Between G1E & G1N
733	3.33 (3)*	2.40 (5)	2.67
267	3.33 (3)	3.00 (2)	2.83
317	3.06 (9)	3.00 (2)	2.17
031	3.04 (10)	-- (1)	3.50
582	2.98 (11)	-- (0)	--
771	2.69 (10)	-- (1)	3.10
332	2.67 (3)	3.00 (2)	2.33
060	2.60 (5)	3.33 (4)	3.31
347	2.39 (9)	2.00 (2)	2.56
773	1.73 (6)	2.00 (3)	2.11
Mean	2.78	2.53	2.70

Note: * Values in parentheses are the number of judges in the group.
-- The number of judges was one or less in the group, therefore comparisons could not be made. They were not included in the mean.

and 2.53. When these means are calculated only across the seven items for which there are data for G1N, the results are 2.73 and 2.68 for G1E and G1N respectively. For five items (733, 267, 031, 317, & 582), the average agreement was 3.00 (rounded) or higher for G1E. For G1N, there were three items on which average agreement could not be calculated. The average agreement for four of the remaining items (267, 317, 332, and 060), was 3.00 or higher. These results suggest an item by group interaction that is further exemplified by the between group agreement data. The between group comparisons showed that when items 031 and 771 were set aside (because there are one or no clinicians in G1N), the agreement reached a value greater than 3.00 only on item 060.

The picture that emerges from the means is that there is greater consistency within G1E than within G1N. When comparing within items means using the 3.00 criterion, the consistency within groups is approximately the same and slightly higher than between group consistency. Where possible, the remaining results will be presented to illustrate the influence of group membership.

Comparison of Clinicians' Judgements Using the NT and Competency Rating Scale

The reason for comparing decisions using the competency rating scale (CRS) approach and the NT procedure of eliminating distractors was to explore the validity of the clinicians' judgements. If clinicians could use two different distractor type methods to eliminate alternatives in a different and reproducible manner, then their judgements for each distractor would be evidence of convergent validity. To achieve this goal, a four-point competency rating scale was created, requiring clinicians to judge the competence of a student who might choose each distractor.

For the CRS ratings, each option was treated as though it was a plausible answer from a student's perspective. Each clinician was asked to estimate the level of competence of a student who chose a specific option by viewing each option as though a student had picked it as his or her answer. In the NT procedure, alternatives are eliminated by clinicians because they think that competently reasoning students should eliminate the option because of its implausibility or inconsistency with the stem. The CRS approach of setting standards compares the option with the stem for correctness. Both methods address each distractor individually but from opposite perspectives, inclusion (CRS approach) versus exclusion (NT procedure) and in this way, contrasting methods to setting standards are provided.

When clinicians rated a distractor on the CRS scale, the ratings could be compared with the NT judgements to see if they were similar. If clinicians eliminated options as implausible, they should also rate them lower on the CRS scale. Thus, the competency rating scale (CRS) decisions served as an audit on the NT judgements.

The CRS scale was designed with four discrete categories, as shown in Table 4.05. Clinicians used only whole numbers to rate the competence of a student who chose each distractor, and zero rating was not allowed.

Table 4.05

The Competency Rating Scale

Rating	Competence Level
1	If a student chose this distractor, I would have serious doubts about their competence.
2	If a student chose this distractor, I would have some doubts about their competence.
3	A student could choose this distractor and still be considered competent.
4	This should be an attractive distractor for a competently thinking student to choose.

Method

To analyze the CRS ratings, the unit of analysis was the clinician's rating for each option within an item. Each clinician's judgement for each distractor was entered on a spread sheet, tallied, and checked. To compare the clinicians' decisions between the NT judgements and CRS ratings, the following procedures were used.

First, either a distractor was eliminated or not eliminated for the NT procedure. Second, the CRS ratings were divided into two groups, those distractors rated 2 or lower,

and those distractors rated 3 or higher. The NT judgements and CRS ratings were tabulated and a 2 X 2 contingency table was created for each item as shown in the example in Table 4.06 for item 317. (The contingency tables for items not specifically

Table 4.06

2 X 2 Contingency Table for Item 317 for G1

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	34	7	41
	3, 4	1	2	3
	Total	35	9	44

discussed in the text are shown in Appendix C.) In Table 4.06, there were 44 distractor judgements in all. Of the 44, there were 35 judgements to eliminate distractors under the NT procedure, and 41 ratings of 1 or 2 using the CRS approach. Convergent validity is illustrated by the consistent judgements. There were 34 judgements both to eliminate and to rate 1 or 2, and two judgements to "not eliminate" and rate 3 or 4. There were 7 + 1 = 8 decisions that were inconsistent.

The clinicians' decisions were examined using two statistics, the phi coefficient and the Chi-Square test. The phi coefficient was used to describe the strength of relationship between the NT and CRS decisions. To determine whether this relationship was significant, the Chi-Square test for 2 X 2 contingency tables was used. Due to the small expected frequencies in two of the four cells for every item, the Chi-Square analysis with a Yates' correction for continuity was used.

Frequencies that occur in the cells are a mixture of different clinicians and

different distractors. Though the independence assumptions of Chi-Square were violated, it was felt that this statistic could be used to provide an approximate indication of the significance of relationship existing between the two methods.

As noted in an earlier section, there is reason to believe that the G1E and G1N groups might approach standard setting differently, so the consistency analysis should have been carried out for the two groups separately. For 7 of the 10 items, the number of G1N clinicians was two or fewer. This meant that the phi coefficient and Chi-Square would have to be calculated on 8 or fewer observations. To get around the problem, the analysis on the combined group (Table 4.07) was done first. Then, the phi coefficient was

Table 4.07

Phi Coefficient and Chi-Square Test with Yates' Correction: NT and CRS Judgements

	G1E		G1		
Item	Phi	Observ.*	Phi	Observ.*	χ^2
317	0.36	36	0.31	44	1.73
267	0.12	12	0.40	20	1.18
771	0.46	40	0.40	44	3.94**
773	0.68	24	0.62	36	11.12**
347	0.65	35	0.67	43	16.61**
031	0.68	40	0.71	44	18.05**
060	0.79	16	0.72	36	14.66**
582	0.76	44	0.76	44	21.74**
332	0.65	12	0.82	20	9.80**
733	0.67	12	0.92	32	22.73**

Note: * Total of clinicians' judgements per item
** Significant at $\alpha = 0.05$ ($df = 1$).

calculated for G1E alone so that any influence of group membership could be detected (Table 4.07). For three the items (267, 332, & 733) there were only three clinicians in

G1E. The phi coefficients for these items are based on only 12 observations and should be treated with caution.

Results of the NT and CRS Comparison

Generally speaking, there was consistency across both methods for most of the items. The phi coefficient showed a nonsignificant relationship between clinicians' NT and CRS decisions only for items 267, 317, and 771 (Table 4.07). The nature of the inconsistency can be found from an examination of their contingency tables (Tables 4.06, 4.08, 4.09, 4.10, 4.11, 4.12).

Of the 44 decisions for distractors for item 317 (Table 4.06) using the NT approach, there were 36 consistent decisions, and 8 inconsistent decisions within G1. Seven of the eight inconsistent decisions occurred with ratings of 1 or 2 on the CRS scale, that is, students who chose the alternative would not be seen as competent, yet according to the NT procedure students might plausibly consider the alternative. These decisions imply diametrically opposed reasoning in that competently reasoning students should consider implausible alternatives. The phi coefficients for G1 and G1E (Table 4.07) of 0.31 and 0.36 suggest that there was little influence of group membership on the relationship. Referring to the verbal protocols, the clinicians explained that irritable bowel disease was a common problem in clinical practice, but students did not have the clinical experience to treat this disease at this point in their medical training.

Clinicians' NT and CRS decisions within G1 also lacked consistency for item 267, as shown in Table 4.08. Content analyses of the clinicians' protocols showed that clinicians were not confident with their choice of the keyed response because many of them were unfamiliar with the discipline of endocrinology. Patients with endocrinological problems were not commonly seen in their daily clinical practices. One subspecialist clinician chose to omit this item because of the lack of contact with patients having

endocrinological problems in clinical practice and an absence of current knowledge in the discipline of endocrinology.

Table 4.08

2 X 2 Contingency Table for Item 267 for G1

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	14	1	15
	3, 4	3	2	5
	Total	17	3	20

The phi coefficients for G1 and G1E for item 267 were 0.40 and 0.12 respectively, suggesting that there was an influence of group membership on consistency. As shown in Tables 4.09 and 4.10, the judgements of the G1N reasoners were very different from the

Table 4.09

2 X 2 Contingency Table for Item 267 for G1E

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	7	1	8
	3, 4	3	1	4
	Total	10	2	12

G1E reasoners. The G1N clinicians' judgements coincided perfectly, that is seven judgements rated 1 or 2 were eliminated, and one judgement rated 3 was not eliminated. This was an instance where converging validity was higher amongst the novice reasoners.

Table 4.10

2 X 2 Contingency Table for Item 267 for G1N

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	7	0	7
	3, 4	0	1	1
	Total	7	1	8

Because of the few judgements, these results should be viewed cautiously.

The results of the analysis of item 771 (Table 4.11) were the most interesting analysis of the 10 items in the study. The Chi-Square test showed that the clinicians' NT

Table 4.11

2 X 2 Contingency Table for Item 771 for G1

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	32	9	41
	3, 4	0	3	3
	Total	32	12	44

and CRS judgements were significantly related, but the phi coefficient (0.40 for G1, 0.46for G1E) showed that the relationship between the two sets of judgements was not strong. From the total of 44 clinicians' judgements, there were 32 NT judgements to eliminate a distractor, with no judgements rated 3 or 4 on the CRS scale. These judgements were in total agreement, exemplifying convergent validity for this item. Of the 12 distractors not eliminated using the NT judgements, nine were rated 1 or 2 on the

CRS scale. These nine judgements imply diametrically opposed reasoning. Students might not eliminate these alternatives (NT procedure) as potential correct responses, but the clinicians would have serious or some doubts of competent student reasoning if they chose these distractors as the correct response (CRS approach).

When the G1E (n = 10) and G1N (n = 1) contingency tables were examined, almost all of the inconsistent decisions occurred in the G1E's (Table 4.12). The clinicians' protocols revealed that the judgements made in this item closely paralleled

Table 4.12

2 X 2 Contingency Table for Item 771 for G1E

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	29	8	37
	3, 4	0	3	3
	Total	29	11	40

daily practice and the management of irritable bowel disease. Maybe due to the differing perspectives of patient management, eight clinicians might have serious or some doubts about a student's competence if they chose the option, but thought that a student might consider the alternative plausible. The G1E's decisions for this item did not converge for the competency ratings of 1 or 2 and not eliminating an option as incorrect.

Item 733, from the discipline of neurological medicine, showed the most consistent decisions of the eight items for G1 (Table 4.07). As shown in Table 4.13, of the 32 judgements, clinicians eliminated 24 distractors using the NT procedure and rated only one distractor 3 or 4. No distractors were rated 1 or 2 and included (not eliminated)

in NT judgements. Nine distractors were eliminated using the NT procedure and were rated 3 or 4, illustrating consistency of judgements. These clinicians' decisions coincided, suggesting that students who chose a distracter rated as 3 or 4 on the CRS scale were reasoning competently.

Table 4.13

2 X 2 Contingency Table for Item 733 for G1

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	23	0	23
	3, 4	1	8	9
	Total	24	8	32

Item 733 was unique because there were more novice clinicians (five) than expert clinicians (three) for this item. The novice clinicians were slightly more convergent than the experts' judgements, as shown in Tables 4.14 and 4.15. This was largely due to the

Table 4.14

2 X 2 Contingency Table for Item 733 for G1E

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	10	0	10
	3, 4	1	1	2
	Total	11	1	12

Table 4.15

2 X 2 Contingency Table for Item 733 for G1N

		Nedelsky Type		
		Eliminate	Not Eliminate	Total
Competency Ratings	1, 2	13	0	13
	3, 4	0	7	7
	Total	13	7	20

experts setting a higher passing score for the items. To explore this result further, the clinicians' protocols were reviewed. Clinicians explained that the knowledge needed to solve this item was factual, detailed, and not commonly used in their daily practice. They said that a student would have this kind of knowledge because they probably had their neuroanatomy course recently. This finding suggests that novice reasoning individuals decisions may coincide with a Phase III student's knowledge level for this type of diagnosis item as compared to an individual classified as an expert. Perhaps the G1N clinicians' knowledge structures were more consistent with the students' reasoning than the G1Es' knowledge structures.

The phi coefficients for the remaining items were significant for the G1 and G1E clinicians (0.65 to 0.80), and the cell frequencies showed little influence of group on consistency. The clinicians' decisions supported the notion of convergent validity.

In summary, the phi coefficient showed a low relationship between the NT judgements and CRS ratings for three items, and showed a strong relationship between the NT and CS ratings for seven items. These findings suggest that when clinicians make judgements of competent student reasoning, using the NT method of excluding distractors, the judgements were frequently and directly related to the ratings used in the

CRS approach.

Where judgements were not significant in two items, the stability of the decisions could be questioned. Clinicians' judgements for these two items, 317 and 267 did not indicate consensus among the clinicians in identifying how competently reasoning students solve items. For item 317, the clinicians' protocols showed that they relied on clinical practice experience to manage this gastrointestinal problem. For item 267, some clinicians voiced concerns about their own lack of knowledge in endocrinology. They explained that the opportunity to apply their basic sciences knowledge obtained in medical school was not available in their clinical practices because of the rarity of patients having these types of illnesses.

It may be the case that clinicians who are asked to set standards of competence peripheral or external to their area of practice could set an inappropriate standard of competent reasoning for students at their level of medical training. The issue of the G1E and G1N differences in judgements showed that inconsistencies occurred within the G1E group and affected the convergent validity of the judgements in these items. However, if clinicians are allowed the opportunity to discuss each item within a multi-discipline committee, these issues may resolve themselves.

Summary of the Judgement Comparisons

The analysis of the clinicians' judgements in Chapter 4 was based on the results of Chapter 3, identifying clinicians who chose the keyed response (G1) and those who did not (G2), and clinicians classified as competent expert (G1E) and novice (G1N) reasoners. Based on these categorizations, the clinicians' judgements of competent student reasoning were explored.

An important issue considered during the analysis was the sequence of the interview and its impact on the judgements made for standard setting using the two distractor procedures. Clinicians thought aloud as they reviewed each item three times after selecting an answer: for standard setting, explaining misconceptions for each distractor, and competency rating. When they set the standard, clinicians were reading the item for the second time. Competency ratings were done the fourth time each clinician read the item. Comparisons between NT judgements and CRS ratings must be viewed cautiously because of the increased familiarity that clinicians had with the item as they made their judgements.

In the first analysis, comparing the passing scores between G1 and G2, the results showed that the mean passing score was higher for G1 than G2. This finding implies that the G1 clinicians would set a higher standard than the G2 clinicians. Also, the G1 clinicians' decisions were slightly more consistent than the G2 clinicians, indicating more stability and reproducibility in their judgements. When the between group agreements were examined, the results sho ed that the groups do differ. Because of these two results, the decision was made to exclude the G2 clinicians from the remainder of the study.

When the passing scores for G1E and G1N were calculated, the results showed that the differences between the two groups were not as clearly defined as for G1 and G2. The results showed the mean consistency of judgements within the G1E clinicians to be slightly higher than the mean consistency for the G1N clinicians. The mean consistency between the G1E and G1N was lower than the mean consistency with G1E. Because of this finding, it was decided that the remaining results would show how group membership (G1E & G1N), affects the validity of the NT and CRS judgements.

In the next section, the auxiliary distractor based standard setting approach, the competency rating, was explained and applied. When the strength of relationship between the two decisions was calculated, a nonsignificant relationship between the decisions was revealed for two items. The relationship between NT judgements and CRS ratings was significant for the remaining eight items, providing evidence in support of the validity of the NT approach to setting standards. The consistency of the clinicians' judgements seemed to rely on three issues: (1) the clinician's expert or novice reasoning ability (position on the knowledge continuum); (2) the clinician's familiarity with the discipline in medicine (procedural knowledge); and, (3) the frequency of the problem arising in clinical practice (application).

Implications of the Clinicians' Perceptions of Competent Reasoning in Students

To set credible standards, clinicians must be competent reasoners and their decisions should be reproducible. If clinicians are not reasoning competently, their perceptions of how competent students reason to select the keyed response may be inaccurate and erroneous. Instability of the clinicians' judgements may result in the setting of unrealistic standards, thereby threatening the validity of their decisions.

The discussion in Chapter 5 is centred on a closer examination of the validity of the NT standard setting process by exploring whether clinicians put themselves in the "minds" of students when they set standards. The analyses focus on two questions. First, when competently reasoning clinicians explain the rationale for eliminating a distractor as incorrect when using the NT procedure for setting standards, do they give reasons that competently reasoning students use? Second, when clinicians explain how a competently reasoning student would think to solve the item, do they use explanations that students use? Competently reasoning students' protocols were used to verify the clinicians' perceptions for both questions. These issues are explored in Chapter 5.

Chapter 5: Clinicians' Perceptions of Students' Reasoning

Introduction

The discussion in Chapter 5 is organized into three major sections and a summary. In section one, the rationale for the analysis is presented. During the study, clinicians set standards by eliminating alternatives. In many cases their think aloud protocols revealed the reasons that they thought students would use to eliminate the alternatives. In section two, comparisons are made between these clinicians' reasons and the reasons that students gave in Skakun's study (1994). In section three, a comparison is made between the explanations that clinicians gave for how students would reason to select the keyed response and the reasons that students in Skakun's study used. Chapter 5 ends with a discussion that integrates the results of these analyses.

Rationale for Analyzing the Clinicians' Perceptions of the Students' Reasoning

Nedelsky (1954) implied that competent judges should examine the item from the perspective of a student. He explained,

> Judging a response in comparison with other responses is theoretically sound, for it probably more closely corresponds to the mental processes of a student. To make a proper judgment of this kind requires time and considerable pedagogical and test-wiseness sophistication" (page 7).

Nedelsky went on to say, "...[T]he difficulty of a test item is a function not only of the question and the form of the correct response but also of the quality of the wrong responses" (p. 11). Both tenets were unexplored in Nedelsky's study, but are central to the discussion in this chapter.

According to Nedelsky (1954), when clinicians examine each item's alternatives, they

should be able to identify which options competently reasoning students should eliminate as obviously incorrect. They should know which options are moderately wrong, an indication of partial knowledge, and which option is the correct response, the keyed option. For clinicians to identify options within these categories, they must be further along the knowledge acquisition continuum than students (Anderson, 1983), but if they are further along the continuum they may not appreciate the complexity and possible ambiguity that would be apparent to competently reasoning students. To determine if clinicians could explain how students solve items-move down the knowledge continuum-two facets needed examination. First, when clinicians were setting the standard using the NT procedure, did they provide a rationale for eliminating a distractor as incorrect that coincided with the students' reasons? Second, did the clinicians' explanations of student reasoning parallel the students' rationales for choosing the keyed alternative? If clinicians successfully performed these two tasks, it would verify their ability to move down the continuum to think from a student's perspective. If the clinicians' perceptions did not parallel the students' responses, then the Nedelsky approach would be called into question.

Method

Since the focus of this inquiry was on the behaviour of clinicians, the analysis centred on whether the reasons clinicians gave agreed with the reasons students gave for eliminating distractors. To determine if the clinicians' insights into the students' reasoning processes were accurate or inaccurate, a content analysis of each protocol was done. An explanation of the procedure used to examine the clinicians' standard setting behaviour and their perceptions of student reasoning follows.

Distractor Analysis. For the distractor analysis, the clinicians' reasons for eliminating

each distractor were categorized and coded. Students' protocols were examined to see if their reasons paralleled the clinicians' reasons. The results of the categorization and coding of the clinicians' reasons for elimination and students' protocols were verified by another rater (Dr. C. Harley). The interrater agreement for the distractor analysis was 80.8% (range 75% to 92%). The verification procedure is described in Appendix D.

Due to the qualitative nature of the analysis of the clinicians' and students' protocols, and due to the spontaneous nature of the clinicians' explanations, the criterion used to refute the Nedelsky theory was set at a fairly low level. When clinicians did not identify more than half of the reasons students used to eliminate each distractor as incorrect the data were taken as not supporting the theory. The proportion of agreement of reasons between the clinicians and students was calculated using the following equation.

$$\text{Proportion of Agreement} = \frac{\text{Number of Reasons Listed by the Clinicians and Used by the Students}}{\text{Number of Reasons the Students Listed by the Clinicians + Other Reasons Listed by the Students}}$$

For example, in Item 347, for alternative two, the clinicians listed one reason that students used. Students used this reason to eliminate alternative two, and listed two additional reasons. The proportion of agreement between the clinicians and students reasons to eliminate alternative two is 1 divided by (1 + 2) = 0.33. According to the criterion of having clinicians identify >50% of the students' reasons to eliminate an alternative, less than half of the reasons listed by the students were identified by the clinicians. The clinicians did not successfully identify the students' reasons for eliminating alternative two.

Skakun (1994) recorded only the final justification for eliminating a distractor. In the few cases where they gave more than one reason, the final one was the dominant factor in their decision.

Reasoning to the Keyed Response. For the keyed response analysis, clinician protocols were examined to determine how a clinician thought a student would choose the keyed alternative. The clinicians' perceptions were categorized and coded. The clinicians' reasons were categorized into four areas: atomistic, holistic, expectations, and no explanation given. An atomistic approach is typified by individuals who solve items by elaborating on specific pieces of clinical information presented in the stem, that is, associating discrete bits of information with each alternative (atomistic). Clinicians providing reasons in the holistic category explained that students would approach item solution by viewing the item from a global perspective. A student would solve the item using concepts, principles, and inductive reasoning (holistic). Other clinicians stated their expectation of a student's knowledge level, and separated their knowledge level from that of a student (expectation). Some clinicians did not explain how a student would reason to solve the item (no explanation).

Students' item solving strategies were compared to the clinicians' explanations. The result of the content analysis was verified by another rater. The interrater agreement for the clinicians' perceptions of how students reasoned was 67% (range 54% to 80%). The interrater agreement for how students reasoned was 69% (range 60% to 77%). The verification procedure is outlined in Appendix E.

Overview of the Chapter

The results in Chapter 4 showed there were differences in reasoning between the G1E and G1N clinicians. This finding imposed certain restrictions on the number of

items that could be analyzed in Chapter 5. Where there was one or no novice clinician reasoners in an item, the item was not used because the results of a an individual's reasoning make a valid intergroup comparison difficult (items 771 & 582). Item 267 was deleted because only three students chose the keyed response. Again, it was felt that there were too few students to validate the clinicians' judgements and perceptions. Because student protocols were needed to verify the clinicians' perceptions of student reasoning, items without student protocols, items' 031, 060, and 773, were not included in Chapter 5's analysis.

The four remaining items are presented in the order of clinician performance, from easiest to most difficult: 347, 317, 733, and 332. These items are presented in the following sequence. First, the discipline, the taxonomy classification, and the number of clinicians and students who chose the keyed response are stated. Then, the numbers of expert and novice clinicians, and students are listed. Second, the clinicians' and students' rationales for eliminating options are compared. Third, the clinicians' perceptions of how students reasoned to solve the item, and the students' item solving behaviours are compared. As shown in Chapter 4, differences in standard setting behaviour existed between experts and novices. These differences will be explained in the qualitative analysis that follows.

The Clinicians' Perceptions of the Students' Competent Reasoning

Item 347

Item 347 was from the discipline of pulmonary medicine and classified as comprehension. Eleven clinicians and 21 of the 40 students selected the keyed option as their response. Eight clinicians and 16 students were classified as experts and three clinicians and five students were novices.

Elimination of Distractors. The analysis of the reasons distractors were eliminated

(Table 5.01) showed that the clinicians' behaviours did not support the Nedelsky theory for alternatives three, four, and five. For alternative one, three expertly reasoning clinicians and one novice clinician stated the three reasons of the students used to eliminate this alternative. Of the three reasons, students used one reason identified by both the expert and novice reasoning clinicians. Students offered an additional reason not stated by either clinician group. The expert clinicians were more successful in identifying the students' reasons for eliminating the distractors. Overall, the clinicians' behaviours supported the Nedelsky theory for alternative one.

Of the students who explained their rationale for eliminating alternative three, one reason was identified by the expert clinicians. None of the three novice reasoning clinicians commented on the issue of risk factors. Students voiced two other reasons to eliminate "Pulmonary embolism". Expertly reasoning clinicians identified one reason the students used, but the combined group of clinicians' judgements refuted the Nedelsky theory for alternative three.

Clinicians classified as expert reasoners identified one reason students used to eliminate distractors for alternative four, and the novice clinicians offered a reason none of the students used. Only the expert reasoning clinicians identified one of the three reasons students used to eliminate alternative four, therefore the clinicians decisions for this distractor refuted the Nedelsky theory. The novice clinicians did not appear to move to the students' levels of reasoning, but the expert clinicians did.

Alternative five posed a problem for both clinicians and students. The expert clinicians listed four reasons, of which students used only one. The novice reasoning clinicians did not state any reasons to eliminate "Pleurodynia". Students stated that they did not know what pleurodynia was. Clinicians overlooked this possibility entirely. As

Table 5.01

Item 347 Summary of Reasons for Eliminating Distractors

Stem: A previously healthy 27 year-old female is suddenly seized with pleuritic pain in the left chest and shortness of breath. The most likely cause is:

Keyed Option: 2. Spontaneous pneumothorax.

Alternative 1. Mycoplasma pneumonia.

Reasons for elimination.	Eliminated by Clinicians E*	Clinicians N**	Students
1. Mycoplasma pneumonia is not characterized as a catastrophic event.	X***	O****	X
2. Additional symptoms are present with mycoplasma pneumonia.	X	X	X
3. The number of pneumonias is unusual.	X	O	X
4. The presentation does not fit with mycoplasma pneumonia.	O	O	X

Alternative 3. Pulmonary embolism.

1. No risk factors are stated.	X	O	X
2. The history does not indicate a pulmonary embolism.	O	O	X
3. The age is inconsistent with pulmonary embolism.	O	O	X

Alternative 4. Acute pericarditis.

1. Suddenly seized is inconsistent with acute pericarditis.	X	O	X
2. Acute pericarditis is quite unlikely due to the absence of symptoms if a viral infection or bacterial infection.	O	X	O
3. Given the framework of this question, this is incorrect.	O	O	X
4. The age is inconsistent with acute pericarditis.	O	O	X

Alternative 5. Pleurodynia.

1. Can be associated with a prodrome or premonitary symptoms, such as a viral infection or fever.	X	O	O
2. Pleurodynia is not associated with a sudden onset of pain.	X	O	X
3. Pleurodynia is not associated with shortness of breath.	X	O	O
4. Pleurodynia can present like this.	X	O	O
5. Do not know what pleurodynia is.	O	O	X

Note: E* = Expert reasoning clinicians; N** = Novice reasoning clinicians.
X*** = Reason given; O**** = No reason given

with alternatives three and four, the clinician decisions for alternative five did not favour the Nedelsky theory.

To summarize the distractor elimination comparisons between the clinicians and students for item 347, the clinicians did not identify the reasons students eliminated distractors successfully for three of the four distractors. The expert clinicians were more successful than the novice clinicians in identifying the reasons students used to eliminate distractors, but the G1E's spontaneous explanations reached criterion only for alternative one. In total, the clinicians did not move down the continuum of knowledge to view the distractor from a student's perspective, therefore these clinicians' justifications for eliminating these alternatives refuted the Nedelsky theory.

Reasoning to the Keyed Response. The clinicians' and students' behaviours and clinicians' perceptions for item 347 are listed in Table 5.02. None of the clinicians thought that the students would use anatomistic approach to solve this item however students did. Student 027, a novice reasoner, who used an atomistic strategy, was typical of most of the novice reasoning students. This student explained:

> *Previously healthy 27 year old female, pleuritic pain left chest, shortness of breath. She's young so its probably not an MI. In any case, pleuritic pain what does that mean? Shortness of breath, what does that mean? Where in the left chest? Is it specific or is it generalized to the chest? Most likely cause is:*
>
> *Mycoplasma pneumonia. Why would you suddenly be seized with pleuritic pain? Spontaneous pneumothorax. Yeah. That is a possibility, but there isn't any history of trauma or anything. Pulmonary embolus is a possibility because of the type of pain, sudden shortness of breath, but previously healthy doesn't make any sense either. Acute pericarditis. That's possible because of the shortness of breath. Would be caused by a disease. She wouldn't want to breathe deep because it would cause her pain and you would get the pleural surface being irritated by the pericarditis. Epidemic pleurodynia. I've never heard of it before but it probably means epidemic pleurodynia. That's a new one.*
>
> *Previously healthy 27 year old. I really can't come up with anything. Spontaneous pneumothorax probably, but that's just a guess among many possibilities. There's more stuff that you want to know for this. You do an x-ray and that's how you make your diagnosis, not from these clinical findings. OK. That one's kind of hazy.*

Table 5.02

Clinicians' Perceptions of and Students' Behaviours for Item 347

		Approach			
Group	Classification	Atomistic	Holistic	Expect.	No Expl.
Clinicians	Novice	O*	X**	O	O
	Expert	O	X	X	X
Students	Novice	X	O	-	-
	Expert	X	X	-	-

Note: O* = No individuals in this category;
X** = Individuals in this category.

When clinicians thought students would use a holistic approach, their perceptions paralleled the students' reasoning. An expert reasoner, Clinician FAM02, explained:

> *I think the word spontaneous implies something sudden and without preceding build up. As a student, I would be looking and saying, "That sounds kind of sudden". As a pneumothorax, I would think, if I remember where it occurs, it should cause pleuritic pain. I can see why it would cause shortness of breath. I think they could think about it that way.*
>
> *When they are taught about spontaneous pneumothorax, it is such a dramatic and classic story that they would likely retain that and recognize this stem as describing a typical situation.*

Student 011, classified as expert, used a similar strategy to solve this item. This student said:

> *A previously healthy (pause) shortness of breath. OK, young girl, pleuritic chest pain, shortness of breath, sounds like pneumothorax in this age group but I will read the answers.*
>
> *Mycoplasma pneumonia. OK, suddenly. No, you wouldn't get that kind of picture. Spontaneous pneumothorax. That's what I said. I'll pick that one first, but I'll read the*

others. Pulmonary embolism. Why would she suddenly have pulmonary embolism at this age? Acute pericarditis. No, I doubt it. Epidemic pleurodynia. ... I doubt it. I'll pick number two.

When clinicians stated their expectation of the level of knowledge students would have, they were explicit. Clinician INT01 (expert reasoner) explained:

They look at the two words. Spontaneous, it wasn't caused by anything, so it jives with previously healthy young person. Secondly, they know a bit about what a pneumothorax is what a pneumothorax causes. Pneumothorax causes pain in the pleura because of the lung collapse. It causes shortness of breath. It all fits. The only other thing they would have to invoke is the probability. The other conditions that are given aren't particularly common either. It is not like it's comparing that to pneumonia. You don't cough with pneumonia. As long as you know mycoplasma pneumonia doesn't present like this typically, they should be able to have some confidence in spontaneous pneumothorax.

When clinicians gave no explanation how students would reason, clinicians explained the symptoms a patient may have. Clinician PUL02 (expert reasoner) said,

It says that the patient is previously healthy, they are young. They have sudden onset of something. Pleuritic chest pain and shortness of breath are the cardinal symptoms for a pneumothorax.

This clinician explained the etiology of the disease, but did not explain how a student would reason to solve this item.

The analysis also revealed that, of the clinicians who thought the students would solve the item holistically, there were both expert and novice reasoners. Students used both atomistic and holistic strategies to solve item 347.

The remaining clinicians either stated their expectations of student reasoning or gave no explanation how students reasoned at all. Overall, clinician perceptions did not coincide with the students protocols.

Item 317

Item 317 was from the discipline of gastrointestinal medicine and classified as management. Eleven clinicians and 32 students chose the keyed response. Nine clinicians and 19 students were classified as expert reasoners. Two clinicians and 13 students were classified as novice reasoners.

Elimination of Distractors. The analysis of the reasons distractors were eliminated (Table 5.03) showed that the clinicians' decisions for the four distractors did not coincide with the students' reasons for eliminating the alternatives. None of five reasons the clinicians cited to eliminate alternative two were used by the students. Students simply stated the scenario did not support lactose intolerance, but did not state specific characteristics of the disease. Content analysis of the clinicians' protocols revealed that clinicians relied on their clinical experience to eliminate "Lactose free". The clinicians did not move down the continuum of knowledge to the students' level. The clinicians' unprompted behaviour did not support the theory that clinicians could interpret how students would reason to eliminate alternative two.

For alternative three, expertly reasoning clinicians stated two reasons the students used to eliminate this distractor, however students supplied five additional justifications that were of a factual nature. Novice reasoning clinicians did not voice any reasons to eliminate alternative three that were voiced by the students. The competently reasoning clinicians' judgements for alternative three refuted the Nedelsky theory.

Of the reasons clinicians voiced to eliminate alternatives four and five, only one reason was used for each alternative. Students listed additional reasons to eliminate the alternatives for each of these two alternatives, low residue and high residue diets. There were differences between the expert and novice clinicians' approaches to explaining the misconceptions inherent in each distractor. The expertly reasoning clinicians were

Table 5.03

Item 317 Summary of Reasons for Eliminating Distractors

Stem: A 24 year-old airline flight attendant complains of feeling tired and losing weight in spite of a good appetite. For the past year she has noticed voluminous, pale, foul-smelling stools. She recalls being told of having bowel difficulty in early childhood and of being fed a diet consisting largely of bananas. Radiological examination discloses an abnormal small bowel follow through. Biochemical analysis of the stool shows an increased amount of fat. The blood picture shows anemia. Which of the following diets would you select for this patient?

Keyed Option 1. Gluten free

Alternative 2. Lactose Free.	Eliminated by Clinicians		Eliminated by Students
Reasons for elimination.	E*	N**	
1. Lactose intolerance does not result in an abnormal small bowel follow though.	X***	X	O****
2. Excess fat in the stool not associated with lactose intolerance.	X	O	O
3. Weight loss not associated with lactose intolerance.	X	O	O
4. Fatigue not associated with lactose intolerance.	X	O	O
5. Anemia not associated with lactose intolerance.	X	O	O
6. This scenario is not a presentation of lactose intolerance.	O	O	X
7. Diarrhea is not present.	O	O	X
Alternative 3. Low Fat.			
1. Not going to help.	X	O	X
2. This is not a malabsorptive diagnosis.	X	O	X
3. No pancreatic or gall bladder disease is present.	O	O	X
4. The mucosa is not damaged.	O	O	X
5. This diet will not address the underlying pathophysiology.	O	O	X
6. Fat is needed in the diet.	O	O	X
7. Do not know what this will do for the problem.	O	O	X
Alternative 4. Low Residue.			
1. Not going to help.	X	O	X
2. Not a large bowel problem.	O	O	X
3. Do not know what this will do for the problem.	O	O	X
Alternative 5. High residue.			
1. Not going to help.	X	O	X
2. Will make their voluminous stools more voluminous.	X	O	O
3. Not a large bowel problem.	O	O	X
4. No diarrhea is present	O	O	X
5. Do not know what this will do for the problem.	O	O	X

Note: E* = Expert reasoning clinicians; N** = Novice reasoning clinicians.
X*** = Reason given; O**** = No reason given

decisive in their decisions. For example, an expert reasoning clinician, PUL02, was particularly succinct by saying,

> *Three, low fat. They should exclude that, because that is not going to help at all. . . . or high residue. I think they should exclude both of those. The lactose-they may consider like I did.*

Clinician PUL01, a novice reasoner, was more reflective when considering the merits of each distractor. This clinician explained:

> *PUL01: I think they should know lactose free is incorrect.*
> *Researcher: Low fat?*
> *PUL01: I would have to think about that. There is so much fat in the stool, they might be able to improve the symptoms by going to a low fat diet. Probably would help.*
> *Researcher: They should entertain it?*
> *PUL01: I think that you would have to entertain it.*
> *Researcher: Low residue? Should they entertain it?*
> *PUL01: Considering the other two options available, there is enough information here that they should not seriously entertain. I think I would take out the residues, both high and low as not being seriously entertainable.*

These two examples illustrate the differences between the reasoning processes of the expert and novice clinicians approaches to explaining the distractor misconceptions. Expert clinicians were decisive and concise in their explanations, whereas the novice clinicians explanations were more reflective and elaborate.

To summarize the clinicians' and students' reasoning for elimination of distractors, the expertly reasoning clinicians' justifications were slightly more consistent with the reasons students used to eliminate the distractors as compared to the novice clinicians justifications. However, neither of the expert nor novice clinicians' dominant perception of reasons students used to eliminate all four distractors coincided with the Nedelsky theory. Neither group of clinicians moved down the continuum to the students' levels of

cognitive development. When the expertly reasoning clinicians' protocols were compared with the novices, there were differences in the decisiveness and brevity of the explanations.

Reasoning to the Keyed Response. The clinicians' perceptions, and students' behaviours are listed in Table 5.04. As shown, no clinicians thought that students would

Table 5.04

Clinicians' Perceptions of and Students' Behaviours for Item 317

		Approach			
Group	Classification	Atomistic	Holistic	Expect.	No Expl.
Clinicians	Novices	O*	O	X**	X
	Experts	O	X	X	X
Students	Novices	X	X	-	-
	Experts	X	X	-	-

Note: O* = No individuals in this category;
X** = Individuals in this category.

use an atomistic strategy to solve this item, but students did use this approach. The strategy used by Student 002, a novice, was typical of the students who solved the item atomistically. This student read the entire item, addressed each option, then selected an alternative:

> *24-year-old flight attendant, so she's young. Feeling tired and losing weight in spite of good appetite, so it's important that her appetite's good even though . . . that she's losing weight, even though she's got a good appetite. For the past year she noted voluminous, pale, foul-smelling stools, to me that seems like a malabsorption problem. She recalls being told of a bowel difficulty in early childhood and being fed a diet consisting largely of bananas. Bowel difficulty in her early childhood, so that so that would suggest that this isn't any kind of infectious cause or something acute, it's something she's been*

having for a long time and especially . . . it's been going on for a year. The diet of bananas I don't know what that would be. Which of the following diets would you select:

I don't know. Gluten free diet might make sense if... she could have celiac disease and so that would make sense for that.

Lactose free diet might also make sense, they could have given bananas either if they didn't want to give milk or if they didn't want to give lactose.

Low fat, no, just cause I don't think it's a problem with . . . it's not a problem with fat I don't think.

Low residue or high residue, I don't think it . . . it doesn't have to do with residue it has to do with malabsorption of something and I can't . . .

I can't decide between one and two, so I am going to guess one. I guess one. Neither of them seems better to me, but I have to pick one.

An expertly reasoning clinician, FAM04, succinctly explained how a student would think through this item using a holistic strategy by explaining:

They would see that she has malabsorption. That is obvious from the voluminous, pale, foul smelling stools and the large amount of faecal fat. What things can give you malabsorption?

The student would say, "I remember that sprue could give you this". Other malabsorption syndromes which wouldn't give her any illness in childhood like small bowel overgrowth, or other illness that have been talked about in gastroenterology or that I have seen on the ward. She has an abnormal small bowel follow through, and is anemic. That would be the most likely diagnosis. That's how I think they would reason.

Student 023, classified as an expert, used a holistic strategy to select "Gluten free". After reading the stem, this student explained:

Well, the first thing I got from the question is malabsorption. She's eating but she's not gaining weight. So she's got some cause of malabsorption. Young female, thinking Crohn's disease, Crohn's disease is a flag that comes up. Now for the past year she's had pale, foul-smelling stools, so she has no history of blood in the past. Foul-smelling suggesting fat in the stool, I think. I don't remember but there's something in the stool I'm looking for. Recently having . . . and then the bowel trouble in previous childhood. Well, the thing that comes up there is gluten enteropathy sprue, so I'm just going to read the answers now and see what they want me to do.

When clinicians stated their expectations of students reasoning, they separated their

knowledge level from that of a student. Clinician PUL02's (expert) expectation was that students would know the clinical presentation of celiac disease. This clinician explained the expectation of how a student would reason through this problem by saying:

> *The patient is young, has a good appetite, so the weight loss is pathologic. They are eating, but not gaining weight. They have abnormal stools, so they are fat malabsorbing, which is a function of small bowel. . . . Yes, they should know that. This history of eating bananas as a child. Gluten is in things like bread and wheat products, so they would do well on bananas. They also do well on bananas if they have a lactose problem as well.*
>
> *You can get abnormal small bowel pathology with gluten or celiac disease. You can get anemia as well because you are not absorbing iron.*

Other clinicians gave no explanation how students would solve this item. For example, a novice reasoning Clinician PUL01 said:

> *This is small bowel disease, a significant disease. You have anemia. You have weight loss. You also have a history that this was present in childhood. This is not a recent onset problem. This has been a lifelong problem. Also, there is radiological abnormality. This is a significant pathological disease. When I think about pathological diseases of the small bowel, you think of maldigestion and malabsorption. Maldigestion relates to pancreatic abnormalities. You can still have a relatively good looking small bowel if you have maldigestion. This involves malabsorption. You are dealing with a small bowel disease. Gluten enteropathy fits. That is the pathology.*

Generally speaking, clinicians' perceptions of student reasoning for item 317 differed from the students' protocols. No clinician said that students would use an atomistic strategy to solve this item, whereas students did. Clinicians' thought that students would use a holistic strategy, and students used this method to solve this item. The remaining clinicians did not show any indication of moving down the continuum, that is, their perceptions of student reasoning did not coincide with the students' strategies. These clinicians either stated their expectation of a student's knowledge level or did not explain how a student would solve this item. These results showed that the clinicians' behaviour

refuted the Nedelsky theory.

Item 733

Item 733 was from the discipline of neurology and classified as diagnosis. Eight clinicians and 18 students chose the keyed response. Three clinicians and nine students were classified as experts. Five clinicians and nine students were classified as novices.

Elimination of Distractors. The results of the analysis of the distractors elimination are shown in Table 5.05. Of the four alternatives, the clinicians' reasons for eliminating the first alternative, "Right cerebral tumour" were inconsistent with the reasons students gave to eliminate this alternative. Students who eliminated this alternative used one of the expert clinicians' reasons, and offered five more reasons not stated by the clinicians. Novice reasoning clinicians did not voice any reasons listed by the students. Clinicians did not move down the continuum to view this alternative from the students' point of view successfully.

The clinicians' and students' rationales for eliminating alternative two, "Trigeminal neuralgia", and alternative three, "Otitis media" were consistent in that reasons stated by the clinicians were used by the students to eliminate these two distractors. Students did not suggest additional reasons to eliminate these two alternatives. The responses to alternative three were unusual because the clinicians identified all of the reasons students voiced to eliminate this distractor. Clinicians effectively moved to the students' level on the continuum for alternatives two and three.

For alternative five, the expert clinicians' stated two reasons used by students to eliminate multiple sclerosis. None of the five novice reasoning clinicians' gave justifications. Students used an additional two reasons not stated by the clinicians. The

Table 5.05

Item 733 Summary of Reasons for Eliminating Distractors

Stem: A 56 year-old man presents with a month history of intermittent right facial pain. On examination he is found to have a diminished right corneal reflex and a slight hearing defect on the same side. The diagnosis is:

Keyed Option. 4. acoustic neuroma.

Alternative 1. Right cerebral tumour.

Reasons for elimination.	Eliminated by Clinician E*	N**	Student
1. Symptoms should be on the opposite side.	X***	X	O****
2. Doesn't explain the pain and diminished corneal reflex	X	O	X
3. Doesn't cause right facial pain.	O	X	O
4. Symptoms suggest cranial nerve involvement, not cerebellum.	O	O	X
5. Symptoms suggest that the problem is not in the right cerebellar area.	O	O	X
5. The symptoms would be more localized.	O	O	X
6. The symptoms would be more generalized.	O	O	X
7. Do not know how this relates to the scenario.	O	O	X

Alternative 2. Trigeminal neuralgia.

1. Physical findings are not totally explained by trigeminal neuralgia.	X	X	X
2. Does not have a diminished corneal reflex.	X	O	O

Alternative 3. Otitis media.

1. Doesn't explain the corneal reflex involvement.	O	X	X
2. Not a probable presentation of otitis media.	X	O	X
3. No prior history of chronic ear inflammation.	X	O	X

Alternative 5. Multiple sclerosis.

1. Multiple sclerosis is more central.	X	O	X
2. Multiple sclerosis does not produce pain.	X	O	O
3. Multiple sclerosis is more specific.	X	O	X
4. Not a typical presentation of multiple sclerosis.	O	O	X
5. Optic nerve involvement is not present.	O	O	X

Note: E* = Expert reasoning clinicians; N** = Novice reasoning clinicians.
X*** = Reason given; O**** = No reason given.

pattern of the clinicians' reasons did not support the Nedelsky theory because the clinicians identified less than half of the reasons students used to eliminate alternative five.

Both groups of clinicians were concise and decisive when they eliminated incorrect alternatives. For example, Clinicians PUL02, classified as an expert, explained:

Researcher: Should a student eliminate right cerebral tumour as not plausible? Should they entertain it or should they cross if off?
PUL02: They should cross that one out.
Researcher: What about trigeminal neuralgia? Should they cross it off or should they entertain it?
PUL02.: They should entertain it because it is a cause of the main complaint, which is [illegible] facial pain. From the physical findings, they should be able to cross that one out.
Researcher: What about otitis media? Should they cross that one out?
PUL02: Yes, I think they should be able to cross that one off.
Researcher: And MS? Should they consider it? Is it a contender?
PUL02: No. With these findings, it would be unusual. It is a more central thing.

INF02, a novice reasoner, was even more precise in saying,

Researcher: Which of these should a student cross off as implausible, as incorrect?
INF02: One, three, and five.
Researcher; One, three, and five. They should entertain two and four?
INF02: Yes.

Little discussion took place when the clinicians' were eliminating alternatives students should know as incorrect. When the clinicians' protocols were reviewed, they explained that the content in this item was mainly recalling knowledge and using one's logic to solve the item.

To recap the distractor analysis for this item, the expert reasoning clinicians explained the students' reasons for eliminating distractors better than the novice reasoners, even though there were more novices than experts. The novices did not move

down the continuum, whereas the experts did. The clinicians' behaviours were contrary to the Nedelsky theory for alternatives one and five, but supported the theory for alternatives two and three.

Reasoning to the Keyed Response. The clinicians' behaviours, perceptions, and student behaviours are listed in Table 5.06.

Table 5.06

Clinicians' Perceptions of and Students' Behaviours for Item 733

		Approach			
Group	Classification	Atomistic	Holistic	Expect.	No Expl.
Clinicians	Novices	X**	X	X	X
	Experts	O*	X	X	X
Students	Novices	X	O	-	-
	Experts	O	X	-	-

Note: O* = No individuals in this category;
X** = Individuals in this category.

Clinician FAM03, a novice reasoner, explained how students would think by saying:

FAM03: They would think, probably about, things stressed as important and curable, of which acoustic neuroma is one of those useful and important diagnoses because you can do something about it. Then they would have remembered something like it usually presents on one side of the face.

They would go back and they would say, "Does everything seem to be on the same side? Yes. Right facial pain. Right corneal reflex. Same side." They would do that pathway routine. They would think something about, does it present quickly or slowly. They may be able to remember that it can be rapid or slow onset. That would be, I think, the process involved. It's a tough one to get through by pure anatomical logic alone.

Researcher: They would have to use more than just logic?

FAM03: Yes. Recall, as well as kind of a reasoning process from the given facts, from

A to B to C. You know A and B because you have heard them before. You can go to C. You see what I mean?

Student 007, an expert reasoner, used a holistic strategy to begin solving this item while reading the stem, and explained:

A 56 year old man presents to his doctor with a month history of intermittent right facial pain. On examination he is found to have a diminished corneal reflex and slight hearing defect on the right. So this tells us that we have probably a lesion involving the 8th nerve, right facial pain is indicative of the 5th nerve, and diminished corneal reflex involves 5 and 7, so this reminds me of a cerebellopontene angle tumour, one of which is acoustic neuroma which I spotted which is one of the five answers.

This student went on to check the remaining options to verify the keyed response, and chose acoustic neuroma as the answer. Other students who used a holistic strategy used a similar line of reasoning to choose the keyed response.

Clinician INT01, a novice reasoner, suggested that students would solve the item using an atomistic strategy by addressing each alternative individually stated:

They would think, a somewhat older male. It has been going on for more than a few days. It is intermittent pain. There are cranial nerves and cells are involved. This is just not referred pain. They would think that there has to be something neurological going on here rather than just referred pain. They would throw out otitis media. They would, from their knowledge of anatomy, throw out right cerebral tumour. It is too hard to explain in terms of that. Trigeminal neuralgia, they would think about. Then, why does he have the hearing loss? They would discard that. They would be left with the two contenders. Probably, because they have heard about acoustic neuroma presenting like this, they would discard multiple sclerosis.

Student 011 (novice reasoner) followed this line of reasoning, as shown from the following excerpt from this student's protocol:

A 56 year old man, defect on the right. Ok. So corneal. Ok. Facial pain, trigeminal neuralgia. Perhaps. Neurological problem definitely. Diminished corneal reflex. Yes. I don't know if it's compression on the cranial nerve 5 and 7. Slight hearing defect as well, cranial nerve 8. Or perhaps, if there is a bone defect down there, that could do something to it. Seeing as how it's on the right side . . . let's read the answers first. Diagnosis is: . . .

and went on to address each option individually, then chose the keyed response, acoustic neuroma. The behaviour Clinician INT01 described was similar to Student 011's reasoning process.

When Clinician PUL02, classified as an expert reasoner, was asked to explain how a student would solve this item, this clinician stated what was expected of a student as distinct from how a student would reason. This clinician said:

I guess they would have to know some of their anatomy of the brain stem, realize that this facial pain is likely a nerve type of pain. It may be from the brain stem. With the hearing deficit, you are looking at cranial nerve eight. You can have other involvement of other brain stem nuclei, so you could affect five as well. You would have to realize that you could get that constellation of symptoms, if they know their neuroanatomy.

Clinician INT02, another expert reasoner, did not view the item from a student's perspective. This clinician explained:

The clues would be the age, the right age. Usually we don't see acoustic neuromas in anybody under the age of 40. The fact that it's intermittent, I don't think helps you a lot. I don't attach any particular significance to it. Intermittency could be a feature of trigeminal neuralgia, which it often is.

The diminished corneal reflex really has to call to mind the sensory pathway from the cornea, and the motor component. The sensory portion is through five. You have also got a hearing defect, that has to be eighth. You have something that produces pain, which is the fifth. You have got a diminished corneal, which involves the fifth, in terms of the sensory component. You have a hearing deficit which involves the eight, so you have got at least two cranial nerves involved. And, where is that possible? It's possible in the brain stem because the fifth and eighth cranial nerve are very close to one another when they come out of the brain stem. The tumour usually develops right at the angle and this is a very common story.

This clinician carefully explained the neuroanatomy, and further on in the interview, this clinician linked the scenario to clinical practice and the preclinical curriculum. No mention was made of how a student would reason to select the keyed response.

The analysis for this subsection of item 733 revealed that clinicians correctly identified how students reasoned to select the keyed response. Novice reasoning clinicians stated that students would use a holistic approach and separated their knowledge from a student's knowledge level. These two strategies are inconsistent with the characteristics of novice reasoners. These results are difficult to interpret because they provided mixed evidence concerning the validity of the Nedelsky theory.

Item 332

Item 332 was from the discipline of pulmonary medicine and classified as diagnosis. Five clinicians and 22 students chose the keyed response. Three clinicians and 14 students were classified as experts. Two clinicians and eight students were novices.

Elimination of Distractors. The results of the analysis in Table 5.07 show that the clinicians' behaviours refuted the Nedelsky theory for all but one alternative. Of the two reasons expert clinicians stated for alternative one, students used one, and offered another one. The novice clinicians' reasons did not offer any reasons that paralleled the students' justifications.

Alternative two was one of three distractors in all four items in which the clinicians voiced all of the reasons students used to eliminate the alternative. The expertly reasoning clinicians successfully moved down the continuum of knowledge to the students' levels to identify the reasons to eliminate this alternative, however the novice reasoning clinicians did not.

For alternative three, "Bronchiolitis obliterans" only one of the expert clinicians'

Table 5.07

Item 332 Summary of Reasons for Eliminating Distractors

Stem: A 28 year-old environmental activist has a history of having had pneumonia four times in the past twenty years. She has had a productive cough "all her life" which is worse in the winter. Physical examination reveals dullness, diminished breath sounds and numerous crepitations below T3 bilaterally. Her fingers are clubbed. The most likely diagnosis is:

Keyed Option. 4. Bronchiectasis.

Alternative 1. Hypogammaglobulinemia.

Reasons for elimination.	Eliminated by Clinicians E	Clinicians N	Students
1. Skin lesion infections are often associated with hypogammaglob.	X	O	O
2. Does not present with these physical findings.	X	O	X
3. Recurrent infections are present with hypogammaglobulinemia.	O	O	X

Alternative 2. Congenital heart disease.

1. Additional findings would be present.	X	O	X
2. Congenital heart disease presents with shortness of breath, cyanosis, and clubbing.	X	O	X
3. History indicative of pulmonary problems.	X	O	X

Alternative 3. Bronchiolitis obliterans.

1. A disease not known by the students.	X	O	X
2. Students could be confused and think that this on one of the manifestations of bronchiolitis obliterans.	O	O	O
3. This is the wrong presentation for bronchiolitis obliterans.	O	O	X
4. This disease is more of an autoimmune problem.	O	O	X
5. Not enough information to diagnose bronchiolitis obliterans.	O	O	X
6. Bronchiolitis obliterans progresses faster.	O	O	X

Alternative 5. Cystic fibrosis.

1. In comparing bronchiectasis with cystic fibrosis, bronchiectasis would be much more likely.	X	O	O
2. Age is inconsistent with cystic fibrosis.	O	O	X
3. Cystic fibrosis affect the upper respiratory tract.	O	O	X
4. More problems are associated cystic fibrosis.	O	O	X
5. Not information to diagnose cystic fibrosis.	O	O	X

Note: E* = Expert reasoning clinicians; N** = Novice reasoning clinicians.
X*** = Reason given; O**** = No reason given.

reasons matched the students' reasons to eliminate this alternative. Students stated four additional justifications not voiced by the clinicians. As with alternatives one and two, novice clinicians did not offer any reasons that were used by the students. Clinicians did not view the misconceptions in this alternative from the students' point of view.

The alternative in which the clinicians' reasons least matched the students' justifications was alternative five, "Cystic fibrosis". The one reason expertly reasoning clinicians offered did not parallel the students' reasons to eliminate this distractor. Students used four other justifications to eliminate this alternative.

When the clinicians' protocols were examined, the expert clinicians' explanations for distractor misconceptions were more elaborate than the novices' explanations. Clinician FAM04, a novice reasoner, explained,

Researcher: Tell me which of these a Phase III students should know is incorrect?
FAM04: None.
Researcher: They should entertain them all?
FAM04: They should entertain them all, just as I did.

There was no further discussion. Clinician PUL02, classified as an expert reasoner, used a different approach.

Researcher: Should they eliminate hypogammaglobulinemia, or should they think about it?
PUL02: They may think about it as a possible answer because you can get the infections with it. Often it will present with skin infections. I think it is much less likely. With straight hypogammaglobulinemia you would not expect all these chest findings and clubbing.
Researcher: Would they think of it as a contender?
PUL02: I don't think so. With the clubbing and the physical findings, they should be able to eliminate that.
Researcher: What about the congenital heart disease?
PUL02: They may think of that because of the clubbing and the current congenital heart disease. If they were thinking that the patient might be in heart failure with these

crepitations, they may think of that. But, if you look back at the history of having repeated pneumonia and a productive cough, they should be looking more at a pulmonary problem.
Researcher: Should students entertain this or should they cross it off?
PUL02: They should be able to cross that off.
Researcher: So far, they have crossed off one and two. And bronchiolitis obliterans? Should they cross that off or think about it? A Phase III student.
PUL02: They may not even know what that is.
Researcher: Do you think they should cross it off if they don't know what it is?
PUL02: Probably.
Researcher: Should they entertain cystic fibrosis or should they cross it off? Is it plausible?
PUL02: It is plausible. If you looked at four and five together, four would be much more likely. They can pick up CF that way.
Researcher: Would students waffle between four and five?
PUL02: No, I think they would cross that off. I think they would be able to come to the right answer.

To summarize the distractor analysis for this item, the expert clinicians stated more reasons students used to eliminate distractors than the novice clinicians. For one alternative (five), none of the reasons the expert clinicians gave to eliminate alternatives was used by the students. Conversely, for alternative two, the expert clinicians listed all the reasons students used to eliminate this distractor. The clinicians' behaviours refuted the Nedelsky theory for three of the four distractors. Clinicians, particularly the novice reasoning clinicians, did not move down the continuum to view the distractors from the students' perspectives.

Reasoning to the Keyed Response. The clinicians' behaviours, perceptions, and students' behaviours for item 332 are listed in Table 5.08. The clinicians classified as expert reasoners stated their expectation of the level of knowledge a student would need to solve the item. Clinician HEG02, classified as an expert reasoner, explained the facts a student would use to solve the problem, but did not explain how a student would think. This clinician explained:

She has had a productive cough all her life. The symptoms are compatible with the findings of bronchiectasis. The cough is worse in the winter, as well. She has had subsequent repetitive episodes of pneumonia which infers to structural damage or immune defect that results in the inability to clear an illness. The clubbing of her fingers and the physical findings is compatible with bronchiectasis.

Table 5.08

Clinicians' Perceptions of and Students' Behaviours for Item 332

		Approach			
Group	Classification	Atomistic	Holistic	Expect.	No Expl.
Clinicians	Novices	X**	X	O	O
	Experts	O*	O	X	O
Students	Novices	X	X	-	-
	Experts	X	X	-	-

Note: O* = No individuals in this category;
X** = Individuals in this category.

Another expert reasoner, Clinician PUL02's expectation of the level of a student's knowledge was brief, and explained that students would solve the item in the following way.

Just by reading the history. A patient with recurrent pneumonia, productive cough, that's quite classic for bronchiectasis. The physical findings can also be consistent with crepitations and clubbing. It all fits with bronchiectasis if they knew anything about that disease.

The expert reasoning clinicians were concise in their explanations of the level of knowledge students needed to solve this item, but did not explain how students would think through the item. These clinicians did not move down the continuum of knowledge to the students level.

The novice reasoning clinicians explained that students would solve the item either holistically or atomistically. Clinician FAM04 (novice reasoner) said that students would use a holistic strategy by stating,

> *They would think recurrent pneumonia, productive cough all her life, which fits bronchiectasis. Clubbing.*

Student 013, an expert reasoner, took this approach by thinking:

> *A 28-year old environmental activist. Her fingers are clubbed. . . . Four times of pneumonia is a bit simplistic. The fact that she's been coughing all her life, worse in the winter-she might have some airway disease, first of all something like asthma. Physical examination reveals some kind of consolidation. However, in the lower lobes [sic] actually suggests airway disease. The fact that she's clubbed suggests a chronic lung disease usually involving some kind of hypoxia.*
>
> *If we go to our answers, the one that initially pops out to me is bronchiectasis, number four, which could actually explain all of this. Bronchiectasis is an airway disease which can cause a chronic cough. It does predispose to pneumonia which would explain what she has now with the dullness and the consolidation in her lower lobes and it is very well known to cause clubbing of the fingers.*

This student used the criteria Clinician FAM04 outlined, recurrent pneumonia, productive chronic cough, and clubbing. Student 013's approach was typical of expertly reasoning students.

Clinician INF01, a novice reasoner, thought that students would use an atomistic approach. This clinician explained:

> *They may choose cystic fibrosis. They may be right. There are extreme variations in patterns. . . . Hypogammaglobulinemia. Low gammaglobulins make a person prone to recurrent infections. They would be thinking about it. It wouldn't be chucked out. . . . They would go on. They would say, "I haven't heard much about this. This is rare." People could be attracted to the familiar. They would put it on the list. Further reasoning would be by looking at the alternatives. Most likely alternatives would be four and five, five better than four. They would be balancing between those two. There is not enough information in a Phase III curriculum and no exposure for them to have seen a*

case of this or to be familiar with the presentation. I wouldn't feel bad if people thought of cystic fibrosis.

Students classified as novice reasoners followed this strategy, by addressing each alternative, and weighing the information in the alternative against the stem. Students tended to focus on each detail of information in the stem. Student 030, an expert reasoner, read the stem and rephrased the information. This student attended to each alternative methodically by thinking aloud:

She probably has hypogammaglobulinemia, which is possible. It could explain her pneumonia's that she's had, four pneumonia's in the past twenty years. Congenital heart disease. Again, you can get an increased number of respiratory infections with congenital heart disease. Bronchiolitis obliterans. I'm not exactly sure what it is so I won't say that one. Bronchiectasis is possible, is very possible actually. Cystic fibrosis in a 28 year old. There's a lot of 28 year olds around with CF but usually they would have much more problems than four pneumonia's. They would have other associated problems. So I think I'm going to go here with bronchiectasis. It sort of goes with the fact that she has had it all her life. It often starts in childhood.

The student novice reasoners tended to solve this item in this manner, however as shown, students classified as experts also used this approach. They didn't verbalize a hypothesis while reading the clinical scenario, but read each option and evaluated its contents in relation to the facts in the stem.

The novice reasoning clinicians successfully identified how the expert and novice students would reason to select the keyed response. The expert reasoning clinicians did separate their knowledge from a student's level. These results suggest that the clinicians might set a standard compatible with the students' knowledge level, but because there were so few competently reasoning clinicians, these results should be interpreted cautiously.

Summary of the Clinicians' and Students' Reasoning Comparisons

The objective of Chapter 5 was to investigate how closely the clinicians' perceptions of students' reasoning coincided with the students' reasoning when clinicians are setting standards. The results reported in Chapter 4 showed that differences were present between the G1E and G1N clinicians' reasoning.

As explained at the beginning of Chapter 5, clinicians were never specifically asked to disclose why a student would reject an alternative. These results of the distractor analysis showed that clinicians do not usually offer a reason to eliminate an alternative as incorrect spontaneously. This clinician behaviour could be attributed to the clinicians' position on the knowledge continuum, that is, the individual's automaticity and efficiency of processing information. Perhaps the expert clinicians' knowledge structures were so well formed that the declarative knowledge was gone and they were using procedural knowledge and composed knowledge only. Where justifications were presented, clinicians did so without prompting.

A summary of the results in Table 5.09 shows the proportion of agreement of reasons for eliminating alternatives between the clinicians and students. As a group, clinicians, did not identify reasons students used to eliminate distractors for 12 of the 16 distractors (75%) for the four items. The expertly reasoning clinicians however were more successful in identifying reasons students used to eliminate distractors as compared the novice clinicians. In a general sense, with both the expert and novice clinicians, neither group effectively viewed the item from a student's levels of knowledge according to the ACT* theory.

Clinicians captured all of the reasons students stated for only three distractors (19%). Clinicians did not identify any of the students' reasons for eliminating one distractor. This pattern of judgements suggests that the clinicians' decisions on which alternatives are

Table 5.09

Summary of the Agreement of Distractors Elimination Between the Clinicians and Students

	Distractor			
Item No.	1st	2nd	3rd	4th
347	0.75	0.33*	0.50*	0.50*
317	0.29*	0.28*	0.33*	0.25*
733	0.17*	1.00	1.00	0.50*
332	0.50*	1.00	0.20*	0.00*

Note: *Clinicians did not successfully identify >0.50 of the reasons students gave for eliminating a distractor.

implausible or plausible may depend on the item's content, the clinicians' domain knowledge, and the clinicians' ability to view the item from a student's perspective. It may be that clinicians are unsure of justifications students use to eliminate alternatives, although this was not specifically probed in the study. Also, it may be the case that if standard setters cannot identify students probable misconceptions, the standard of performance might be set at a level inconsistent with the students' knowledge levels at this point in their medical training.

As shown in Table 5.10, the results of the clinicians' perceptions of how students reasoned to solve the four items varied. The novice reasoning clinicians thought students would use a variety of strategies to reason to the keyed response. Novice clinicians gave more explanations that were parallel to students' solutions whereas expert clinicians described students' solutions in more general terms (i.e., expectations). This finding suggests that novice clinicians were more likely to move down the knowledge continuum to the students position on the ACT* continuum. This group of clinicians' judgements,

the competent novice reasoning clinicians, would be in concert with the students cognitive development.

Table 5.10

Summary of the Clinicians' Perceptions of the Students' Reasoning to Solve Items

	Novice Clinicians				Expert Clinicians			
Item	A	H	E	NE	A	H	E	NE
347	O*	X**	O	O	O	X	X	X
317	O	O	X	X	O	X	X	X
733	X	X	X	O	O	X	X	X
332	X	X	O	O	O	O	X	O

Note: O* = No clinicians;
X** = Clinicians in this category.

None of the expertly reasoning clinicians said that students would solve any of the items using an atomistic approach. These clinicians thought that students would use a higher level of knowledge to reason to the keyed response, that is, students would use a conceptual, principled approach to solving the item. The expert reasoning clinicians tended to state their expectation of the students' knowledge levels, and frequently did not explain how students would reason to select the keyed response. This group of competently reasoning clinicians did not move to the students positions on the ACT* continuum. Clinicians in this category might set standards of performance beyond the students capabilities. The standard may reflect an unrealistically high passing score.

The findings of the clinicians' perceptions of students' reasoning for solving items revealed two prominent issues. First, expertly reasoning clinicians were more successful in identifying the students' justifications for eliminating distractors than novice reasoning

clinicians. There were fewer novice reasoning clinicians in this study, therefore a note of caution in interpreting the results is needed. Second, novice reasoning clinicians were more successful than expert reasoning clinicians in explaining how students reasoned to select the keyed response. These results suggest that the competently reasoning expert and novice clinicians' decisions complimented each other in identifying the distractor elimination and perceptions of how students solved items. Both groups of clinicians, in a committee setting, might set standards of performance commensurate with the students levels of development at this point in the students medical training.

Chapter 6 - The Findings, Implications, and Future Research

Introduction

Chapter 6 is divided into two sections. The first brief section introduces three areas of the results that form the body of the discussion. Section two addresses each of the three areas from three perspectives. This chapter closes with final remarks on standard setting procedures and comments on the implications of this study to education in general.

Organization of the Chapter

The research question that prompted this study was "How valid are the judgements in standard setting, using a distractor approach, when placed in a cognitive psychology framework?" Examining this question required both qualitative and quantitative analyses to describe the clinicians' reasoning and judgements in solving items and setting standards. The findings fell into three broad areas: (1) standard setting using the NT approach, (2) standard setting within the cognitive psychology framework, and (3) measurement issues related to setting standards. Each area is analyzed from three perspectives: (1) the findings, (2) implications of the findings, and (3) suggestions for future research where appropriate.

Standard Setting Using the Nedelsky Type Approach

As explained in Chapter 5, Nedelsky (1954) implied that competent judges should examine the item from the perspective of a student. According to Nedelsky, when clinicians examine each item's alternatives, they should be able to identify which options

competently reasoning students should eliminate as incorrect. They should know which options are moderately wrong, and which option is the correct response, the keyed option. Although Nedelsky did not provide a cognitive psychology framework for his beliefs, his notion of the reasoning process used by students to solve multiple choice items was validated by Skakun (1994) and Skakun et al. (1994) who found that students often solved items by addressing each alternative individually. In other work on standard setting, however, Triska et al. (1995) found that clinicians had difficulty viewing an item from a student's perspective.

Clinicians who participated in the present study were professional medical educators and considered experts in their area of specialization. Because clinicians solved each item before setting the standard of performance, they became aware of the problem space students need to generate in order to solve the item. Clinicians were sensitized to the alternatives, that is, which alternatives students at a particular level should know are incorrect. The assumption, according to the Nedelsky theory was that clinicians would view the item from a student's perspective and know which distractors students should eliminate as incorrect.

According to the findings in this study, the Nedelsky theory should be revised to place it within an information processing framework such as the ACT* theory. By placing the students and clinicians problem solving behaviours on a continuum of knowledge (Anderson, 1993), the transition from novice to expert reasoner can be explained. In this study, clinician and student novice reasoners were separated out on a continuum from the expert reasoning clinicians and students. Moreover, individuals who did not select the keyed response did not seem to have sufficient declarative knowledge to solve the item. They formed a third group that was further down the continuum than the competently reasoning individuals. The fact that there appeared to be three groups of clinicians

complicated the analysis and influences the discussion that follows. The remainder of this section is devoted to the findings related to setting standards in this study, implications of these findings, and research opportunities related to each finding.

1. **Finding: Clinicians NT judgements were reproducible using another distractor approach (CRS), thereby supporting the validity of the NT procedure.**

   Implications: The CRS approach was taken as an alternative to the NT approach and since there was agreement between the two, the finding was construed as supporting the validity of the NT approach. Equally plausible is that the CRS approach is a useful strategy in its own right. It may be the case that setting standards using the CRS approach could be better than the Nedelsky procedure. Using the CRS approach, clinicians viewed distractors from an inclusion perspective, that is, estimating the competence of a student who chose the alternative, rather than requiring judges to use an exclusion perspective, that is identify alternatives that students should exclude. Although no direct data were collected on the issue, clinicians commented that the CRS was easier to do.

   A factor to consider when using the CRS scale in determining competent reasoning is how should a passing score be calculated? This was not developed in this study.

   Research: Three tasks need addressing: (1) explore the appropriate number of categories; (2) translate the ratings into standards, and (3) calculate the passing score.

2. **Finding: G2 clinicians (those who did not choose the keyed alternative initially) set lower passing scores on each item as compared to the G1 clinicians (those who chose the keyed alternative).**

   Implication: Lower passing scores permit individuals who do not have sufficient domain knowledge to practice medicine. This may result in a passing score set at an inappropriate level with the credibility of the scores becoming questionable. In high stakes testing in the professions (i.e., physicians, lawyers, chartered accountants) the certification standards and licensing organizations' credibility may be threatened.

   Research: This finding should be explored in the context of a practical standard setting situation to determine if training and group discussion can mitigate the problem.

3. **Finding: G2 clinicians judgements had a lower average consistency than the G1 clinicians judgements.**

   Implication: G2 clinicians may not have had sufficient domain knowledge to select the keyed response and were unsure of which alternatives to eliminate as incorrect. These clinicians' judgements could affect the stability of the passing score or lower the cut-off score.

   Research: As noted in finding number two, comparisons of passing score decisions should be done within and between the G1 and G2 clinicians in a practical setting to determine if significant differences are present.

4. **Finding: The G1N clinicians' judgements had a slightly lower average consistency than G1E clinicians' judgements.**

   Implication: The novice reasoning clinicians knowledge structures were not as well

structured as the expertly reasoning clinicians. With the experts being further up the continuum from the novices, the novices' behaviours showed that they did not have the automaticity, efficiency, and confidence of the experts.

Research: If this result occurs in the practical situation, the influence of expertise on the passing score calculated in the practical setting should be investigated. Possible moderating effects of committee discussion should also be examined.

Standard Setting Within a Cognitive Psychology Framework

As previously explained, the focus of the present study was to examine the validity of clinicians' judgements in setting standards in a cognitive psychology framework. In standard setting, competence and expertise are defined with respect to a small piece of a field-that represented by an item. In general research on clinical problem solving, competence and expertise is taken to be a more general characteristic of individuals.

Competence in this study was narrowly defined as selecting the keyed response for each item. Expertise, a subcategory of competence, was determined by evaluating the reasoning process used by each clinician and student as they solved each item. Three levels of knowledge were present, incompetent reasoners, competent novices and competent experts. This study focussed on the competent reasoners, the expert and novice reasoning clinicians and students.

1. **Finding: Three groups of judges were found (experts, novices, and those who did not select the keyed response).**

   Implication: There was little distance between clinicians and students who did not choose the keyed response. The novice reasoning clinicians and the novice reasoning students were in close proximity on the continuum. Expert reasoning clinicians and

expert reasoning students displayed similar characteristics. For standard setting judgements to be valid, clinicians should be further removed, up the continuum from the students. Clinicians should have a domain of well-structured knowledge to determine the level of knowledge students need to practice.

Research: More research is needed to understand how the general characteristics of competence and expertise become translated into specifics at the item level.

The findings should be replicated to determine if significant differences are present in other populations (e.g., lawyers, accountants, nurses).

2. **Finding: Competence and expertise should not be used interchangeably. They are item dependent and should be clarified when they are used.**

Implication: Various researchers have implied that competence and expertise are synonymous, and when an individual meets a certification and licensing requirements, that the individual is competent and expert in their area of specialization. This suggests that individuals who participate in standard setting activities are competent and experts in their area of specialization. The cognitive competence of an individual should be determined before standard setting procedures are initiated. This procedure would assess the individual's domain knowledge (competence) and determine the level of cognitive development (novice or expert). Without this preliminary assessment of judges, the validity of the judgements may be difficult to defend.

Research: The study should be replicated with a larger number of clinicians to determine if significant differences are present.

3. **Finding: The calibre of reasoning was content dependent.**

   Implication: Each individual's knowledge domain differs due to the structure and development of procedural knowledge and practice. Because of these individual differences, clinicians' experience played a vital role in explaining misconceptions in distractors.

   Research: The effect of discussion on setting standards within a heterogeneous group (family physicians & specialists) should be investigated.

4. **Finding: Clinicians more often explained how students reasoned to solve a diagnosis item than how they reasoned for management and comprehension items.**

   Implication: The knowledge needed to solve these diagnoses items was factual, detailed, and hierarchically structured. The clinicians' domain declarative knowledge was transformed into procedural knowledge though automaticity, efficiency, and practice for the diagnosis items. For the management and comprehension items, clinicians were accessing procedural knowledge and using their practice to describe student reasoning. Less declarative knowledge was used.

   Research: The effect of clinician training in cognitive psychology (the students' position on the knowledge continuum) should be compared by examining the item statistics before and after training on items classified as diagnosis, management, and comprehension.

Measurement Issues Related to Standard Setting

The focus of the present study was to examine the validity of clinicians' judgements

in setting standards in a cognitive psychology framework. Even though the topic was narrow, measurement issues arose from the results. A discussion of these findings follows.

1. **Finding: Clinicians did not spontaneously separate their knowledge level from a student.**

   Implication: Not all items may have been created at an inappropriate student's level of knowledge, therefore an unrealistic sampling within a knowledge domain occurred. Because experts appeared to have had greater difficulty in explaining the students' reasoning processes, novice reasoners must be included in the item writing process.

   Research: The effect of cognitive psychology training and item writing for clinicians on item analysis and distractor analysis should be investigated. Distractor analyses should be done to compare the effect of clinicians responses before and after prompting to explain misconception in distractors, then compared with reasons students gave to eliminate distractors.

2. **Finding: Clinicians may have had difficulty in moving down the continuum of knowledge to view the item from a student's perspective.**

   Implication: If clinicians can not view the item from a student's perspective, an inappropriate domain sampling may occur. Items may be created an level inconsistent with the students' knowledge structures.

   Research: Item statistics should be compared before and after clinicians were given instruction on the students' position on the knowledge continuum. Item statistics should be compared before and after clinicians were prompted to explain how students' reasoned to select the keyed response.

Recommendations for Standard Setting Procedures

The findings of this study served to recommend three guidelines for setting valid and reproducible standards of performance. First, competence and expertise are not synonymous with certification and licensure. Before individuals participate in standard setting activities, each committee member's level of competence should be assessed to determine if sufficient domain knowledge is present. The criteria for competent reasoning should be identical for clinicians and students-the individual should solve the item and choose the keyed response. When competent reasoning is established, the level of development (novice or expert) should be evaluated using the six dimensions proposed by Royer et al. (1993).

Second, judges require an understanding of certain aspects of cognition, particularly knowledge acquisition and the position of clinicians and students on the knowledge continuum. Clinicians must be made aware of their level of automaticity, efficiency of processing information, and highly proceduralized and composed knowledge structures, and how they differ from the students. Without this awareness, clinicians may set unrealistic standards for students at a particular level of training.

Third, standard setting committee members should be aware of the role of the experts' and novices' knowledge structures and decision-making characteristics to use both effectively. This study showed that experts' judgements had higher average consistency than the novices' judgements, however the novices were sometimes more successful in viewing an item from a student's perspective. Having both experts and novices reach consensus supports the reproducibility and validity of the standard setting judgements.

Closing Remarks

To set fair, defensible, and accurate standards, the underlying cognitive structures on which the judgements are made, must be solid. This fundamental notion seems reasonable, however it was neglected in researchers' agendas for the past four decades. Historically, psychometricians focussed their research agendas on the technical facets of setting standards of performance. These issues were explored at length, but the psychological constructs that judges used to make their decisions were only alluded to in their articles. Integral to the standard setting process is the assessment of competent thought in individuals. The time has come to integrate cognitive psychology with psychometrics in order to produce valid and reproducible standard setting judgements of performance.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist, 48*(1), 35-44.

Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). New York: W. H. Freeman.

Andrew, B. J. & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement (36)* 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.

Ayton, P. (1992). On the competence and incompetence of experts. In G. Wright & F. Bolger (Eds.). *Expertise and decision support* (pp. 77-102). New York: Plenum.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4*(2), 219-240.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*(2), 145-163.

Chi, M. H. T., Glaser, R., & Rees, E. (1981). *Expertise in problem solving.* Pittsburg, PA.: Learning Research and Development Center.

Cizek, G. J. (1991). *An investigation into one alternative to the group-process procedure for setting performance standards on a medical specialty examination.* Unpublished doctoral dissertation, Michigan State University, Michigan.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cramer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgement. In K. A. Ericsson & J. Smith (Eds.). *Toward a General Theory of Expertise*, (pp. 195-217). New York: Cambridge University Press.

Cross, H. L., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher examinations. *Journal of Educational Measurment, 21,* 137-146.

Doolittle, P. E., & Yekovich, F. R. (1994). Developing expertise in the professions: Theoretical and practical concerns. *Professions Education Researcher Quarterly, 15*(2), 1-5.

Ebel, R. L. (1972). *Essentials of Educational Measurement.* Englewood Cliffs, NJ: Prentice Hall.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical problem solving: A ten year perspective. *Evaluation & Health Professions, 13*(1), 5-36.

Emrick, J. A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement, 8,* 321-326.

Fitzpatrick, A. R. (1989). Social influences in standard setting: The effects of social interaction on group judgements. *Review of Educational Research, 59*(3), 315-328.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519-521.

Glaser, R. (1989). Learning theory and the study of instruction. *Annual Review of Psychology, 40,* 631-666.

Glaser, R., Lesgold, A., & Lajoie, S. (1985). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. Glover, J. C. Conoley, & J. C. Witt (Eds.). *The influence of psychology on testing and measurement* (pp. 41-85). Hilldale, NJ: Lawrence Erlbaum Associates Publishers.

Groen, G. J. & Patel, V. L. (1989). The relationship between comprehension and reasoning in medical expertise. In M. Chi, R. Glaser, & M. Farr (Eds.). *The nature of expertise* (pp. 287-310). Hilldale NJ: Lawrence Erlbaum.

Haladyna, T. M. (1994a). A research agends for licensing and certification testing validation studies. *Evaluation & The Health Professions, 17*(2), 242-256.

Haladyna, T. M. (1994b). *Developing and validating multiple-choice items.* Hilldale NJ: Lawrence Erlbaum.

Hambleton, R. K., & Eignor, D. R. (1980). Minimums, competency testing and social policy. In R. M. Jaeger & C. Kehr Tittle (Eds.). *Minimum competency achievement testing: Motives, models, measures, and consequences* (pp. 367-396). Berkeley, CA: McCutchan.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integrated theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*(3), 159-170.

Hambleton, R. K., & Powell, S. (1983). A framework for reviewing the process of standard setting. *Evaluation & The Health Professions, 6*(1), 3-24.

Harasym, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. *Educational and Pychological Measurement, 41*, 725-734.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.). *Educational measurement* (3rd. ed.) (pp. 485-514). New York: MacMillian.

Johnson, E. J. (1989). Expertise and decision under uncertainty: Performance and process. In M. Chi, R. Glaser, & M. Farr (Eds.). *The nature of expertise* (pp. 209-228). Hilldale NJ: Lawrence Erlbaum.

Kane, M. T. (1992). The assessment of professional competence. *Evaluation & The Health Professions, 15*(2), 163-182.

Kane, M. T. (1994). Validating interpretive arguements for licensure and certification examinations. *Evaluation & The Health Professions, 17*(2), 133-160.

Kennedy, M. M. (1987). Inexact sciences: Professional education and the development of expertise. *Review of Research in Education, 14,* 133-167.

Keren, G. (1992). Improving decisions and judgements: The desirable versus the feasible. In G. Wright, & F. Bolger (Eds.). *Expertise and decision support* (pp. 25-44). New York: Plenum.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement, 17*(3), 167-178.

Kriewall, T. E. (1972). *Aspects and applications of criterion-referenced tests* (IER Tech. Paper No. 103) Dowers Grove, IL: Institute for Educational Research.

LaDuca, A. (1980). The structure of competence in health professions. *Evaluation & The Health Professions, 3*(3), 253-288.

Lemieux, M., & Bordage, G. (1992). Propositional versus structural semantic analysis of medical diagnostic thinking. *Cognitive Science, 16*, 185-204.

Lesgold, A. M. (1983). *Acquiring expertise* (UPITT/LRDC/ONR/ PDS-5). Arlington, VA: Personnel and Training Research Programs.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance of educational and occupational tests*. Princeton, NJ: Educational Testing Service.

McGaghie, W. C. (1980). The evaluation of competence. *Evaluation & The Health Professions, 3*(3), 289-320.

McGaghie, W. C. (1991). Professional competence evaluation. *Educational Researcher, 20*(1), 3-9.

McGaghie, W. C. (1993). Evaluating competence for professional practice. In L. Curry & J. F. Wergin (Eds.). *Educating professionals* (pp. 229-261). San Francisco CA: Jossey-Bass.

Maguire, T. O., Skakun, E., & Harley, C. (1992). Setting standards for multiple choice items in clinical reasoning. *Evaluation & The Health Professions, 15*(4), 434-452.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. *Review of Educational Research, 46,* 133-158.

Meskauskas, J. A. (1986). Setting standards for credentialling examinations. *Evaluation & The Health Professions, 9*(2), 187-203.

Millman, J. (1972). *Tables for determining number of items needed on domain-referenced tests and number of students to be tested* (Technical Paper No. 5). Los Angeles: Instructional Objective Exchange. April.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research 43,* (pp. 205-216). (ERIC Document Reproduction Service No. ED 065 555, 17pp.)

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. *Journal of Educational Measurement, 20,* 283-293.

Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.). *Licensure testing: Purposes, procedures, and practice* (pp. 219-252). Lincoln, NE: Buros Institute of Mental Measurments.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational & Psychological Measurement, 14,* 3-19.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24*(1), 56-64.

Novick, M. R., & Lewis, C. (1974). Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Poppham (Eds.). *Problems in Criterion-Referenced Measurement* (CSE Monograph Series in Evaluation, No. 3, pp. 139-158). Los Angeles: Centre for the Study of Evaluation, University of California.

Novick, M. R., Lewis, C., & Jackson, P. H. (1973). The estimation of proportion in groups. *Psychometrica, 38,* 19-45.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solutions strategies in medical reasoning. *Cognitive Science, 10,* 91-116.

Patel, V. L., & Groen, G. J. (1991a). The general and specific nature of medical expertise: A critical look. In K. A. Ericsson & J. Smith, (Eds.). *Toward a general theory of expertise* (pp. 93-125). New York: Cambridge University Press.

Patel, V. L., & Groen, J. (1991b). Developmental accounts of the transition from medical student to doctor: Some problems and suggestions. *Medical Education, 25*(6), 527-535.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard setting. *Educational Measurement: Issues and Practice, 10*(2), 15-16, 22, 25-26.

Reimann, P., & Chi, M. T. H. (1989). Human Expertise. In K. J. Gilhooly (Ed.). *Human and machine problem solving* (pp. 161-191). New York: Plenum.

Roudabush, G. E. (1974). *Models for a beginning theory of criterion-referenced tests.* Presented at the meeting of the National Council on Measurement in Education, Chicago.

Royer, J. M., Cicero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*(2), 201-243.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine, 65*(10), 611-621.

Shanteau, J. (1992). The psychology of experts: An alternative view. In G. Wright & F. Bolger (Eds.). *Expertise and decision support* (pp. 11-21). New York: Plenum.

Shepard, L. A. (1980). Technical issues in minimum competency testing. In D. C. Berliner (Ed.). *Review of education in research* (pp. 30-84). Washington DC; American Educational Research Association.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.). *A Guide to Criterion-Referenced Test Construction* (pp. 169-198). Baltimore: John Hopkins University Press.

Skakun, E. N (1994). *Strategy choices in responding to multiple items.* Unpublished doctoral dissertation. University of Alberta, Edmonton, Alberta, Canada.

Skakun, E. N., & Kling, S. (1980). Comparability of methods of setting standards. *Journal of Educational Measurement, 17*(3), 229-235.

Skakun, E. N., Maguire, T. O. & Cook, D. A. (1994). Strategy choices in responding to multiple choice items. *Academic Medicine, October Supplement, 69*(10), 57-59.

Tamblyn, R. (1994). Is the public being protected? Prevention of suboptimal medical practice through training programs and credentialing examinations. *Evaluation & The Health Professions, 17*(2), 198-221.

Triska, O. H., Skakun, E. N., Maguire, T. O., & Harley, C. (1995) Judges' standard setting behaviour and students' thought processes in multiple choice examinations. In A. I. Rothman & R. Cohen (Eds.), *The Proceedings from The Sixth Ottawa Conference on Medical Education, Toronto, ON.* (pp. 227-231). Toronto, ON: University of Toronto Bookstore Custom Publishing.

Appendix A: Multiple Choice Item Used in the Study

ITEM ID. 347
A previously healthy 27 year-old female is suddenly seized with pleuritic pain in the left chest and shortness of breath. The most likely cause is:

1. mycoplasma pneumonia.
* 2. spontaneous pneumothorax.
3. pulmonary embolism.
4. acute pericarditis.
5. pleurodynia.

ITEM ID. 332
A 28 year-old environmental activist has a history of having had pneumonia four times in the past twenty years. She has had a productive cough "all her life" which is worse in the winter. Physical examination reveals dullness, diminished breath sounds and numerous crepitations below T3 bilaterally. Her fingers are clubbed. The most likely diagnosis is:

1. hypogammaglobulinemia.
2. congenital heart disease.
3. bronchiolitis obliterans.
* 4. bronchiectasis.
5. cystic fibrosis.

ITEM ID. 733
A 56 year-old man presents with a month history of intermittent right facial pain. On examination he is found to have a diminished right corneal reflex and a slight hearing defect on the same side. The diagnosis is:

1. right cerebral tumor.
2. trigeminal neuralgia.
3. otitis media.
* 4. acoustic neuroma.
5. multiple sclerosis.

ITEM ID. 267
The investigation of bilateral gynecomastia in a 37 year-old male should involve determining all of the following, **EXCEPT**:

1. the serum testosterone.
2. a history of marijuana usage.
3. the serum chorionic gonadotropin level (hCG).
4. the patient's thyroid status.
* 5. ultrasound of the abdomen.

ITEM ID. 60
Iron deficiency anemia is associated with all of the following, **EXCEPT**:

1. an increased total iron binding capacity of the serum.
2. a decreased serum ferritin.
* 3. the presence of ringed sideroblasts in the bone marrow.
4. a low reticulocyte count.
5. erythroid hyperplasia in the bone marrow.

ITEM ID. 582
One week after an anterior myocardial infarction, a 55 year-old man complains of severe pain in the left leg. The leg is cool, pale, and pulseless. The most likely diagnosis is:

1. deep venous thrombosis.
2. ruptured left iliac aneurysm.
* 3. arterial embolism.
4. A-V fistula.
5. arterial thrombosis.

ITEM ID. 773
The most appropriate topical agent in the treatment of contact dermatitis is:

1. calamine lotion.
2. vioform-hydrocortisone comb.
* 3. corticosteroid.
4. antibiotic-corticosteroid comb.
5. colloidal oatmeal bath.

ITEM ID. 317
A 24 year-old airline flight attendant complains of feeling tired and losing weight in spite of a good appetite. For the past year she has noticed voluminous, pale, foul-smelling stools. She recalls being told of having bowel difficulty in early childhood and of being fed a diet largely consisting of bananas. Radiological examination discloses an abnormal small bowel follow through. Biochemical analysis of the stool shows an increased amount of fat. The blood picture shows anemia. Which of the following diets would you select for this patient?

* 1. Gluten free
2. Lactose free
3. Low fat
4. Low residue
5. High residue

ITEM ID. 771
The irritable bowel syndrome in adults is a diagnosis of exclusion. However, when this diagnosis is finally made, you should:

1. tell the patient the symptoms are always due to emotional stress.
2. tell the patient to take tranquilizers when symptoms flare.
3. tell the patient to return for a complete reevaluation (x-rays, blood work, etc.,) in three months.
* 4. counsel the patient and prescribe metamucil and bran.
5. counsel the patient and prescribe Lomotil and Kaopectate.

ITEM ID. 31
An 82 year-old man has neck, hip, and knee pain with joint swelling. The rheumatoid factor test is weakly positive at 1:80. The most likely diagnosis is:

1. late onset rheumatoid arthritis.
* 2. osteoarthritis.
3. polyarteritis nodosa.
4. polymyalgia rheumatica.
5. metastatic malignancy.

Appendix B: Clinicians' Interview Package

CONSENT FORM FOR PARTICIPANTS

Researcher: Olive H. Triska, PhD Candidate, Faculty of Education, Educational Psychology
Centre for Research and Applied Measurement in Education,
3-102 Education North, University of Alberta,
Edmonton, Alberta, T6G 2G5
Telephone: 492-3762, FAX: 492-3179

I understand that I am volunteering to participate in a project in which I will be asked to describe my thoughts while assessing the level of clinical reasoning needed to respond to multiple choice questions in medical examinations. I am willing to share my thoughts with the researcher while "thinking aloud" as I review and answer multiple choice questions of the level of difficulty of fourth year medical students. I understand that the interview will take about two and one half hours and will be audiotaped.

I am aware that the purpose of this research is to examine competent clinical reasoning in medical education multiple choice items from the clinicians' perspective. I understand that my participation in the study will probably not be of direct benefit to me but the results of the study may contribute to the greater understanding of the cognitive processes that influence notions of competent reasoning in medical education.

The researcher has explained this project to me and I have had the opportunity to ask questions about the study. I understand that my identity will remain anonymous in any reports concerning the project and the audiotapes will be kept confidential and secure. I will be given the opportunity to withdraw from the study at any time without prejudice.

I am satisfied that I have been given sufficient information about the study and I am willing to participate in this project by sharing my thoughts on competent reasoning skills.

_____ _____

Date Participant

_____ _____

Date Researcher

COMPONENTS OF A MULTIPLE CHOICE ITEM

STEM

The most definitive diagnostic test for pulmonary embolism with or without infarction is:

ALTERNATIVES

Distracters/Incorrect Alternatives/Incorrect Options

1. Perfusion lung scan
2. Ventilation - perfusion lung scan
3. A decreased arterial pCO_2
4. An increased alveolar-arterial oxygen difference

Keyed Answer/Correct Answer/Correct Alternative/Correct Option

5. Pulmonary angiogram

INTERVIEW PROCEDURE

1. I will be asking you to think aloud as you read each item and select the correct response from the options. I am doing this to help you see the item as the student sees it. After you have selected your answer, I will give you the keyed answer for the item.
2. I will ask you to examine each distracter in turn and indicate the distracters that students should be able to eliminate as incorrect, leaving those alternatives that a competently thinking Phase III student might consider plausible.
3. Then, I will ask you to answer some standard setting questions.

As you work your way through each task, I will ask you to think aloud so that I can record your reasoning processes. By think aloud, I mean: tell me what you are thinking when the thoughts are formed, as you read each sentence in the item. Do not try to reformulate or paraphrase the thoughts before you speak. It is important to continue to tell me your thoughts as they occur. Please try to tell me everything that goes through your mind, even if it seems trivial or tangential to the task.

PART 1

Practice Item

Read the item and tell me what you think the answer is. As you do this, THINK ALOUD so that I can understand how you arrived at your decision. When you are finished, I will give you the keyed answer.

In patients with non-penetrating cardiac trauma, the most common cardiac injury is laceration or rupture of the:

1. AV valve
2. Semilunar valve
3. Coronary artery
4. Left ventricle
5. Papillary muscle

The answer provided by the physician who composed this item is alternative 5, the papillary muscle.

PART 2

Let's go on to and do the standard setting for this item. Remember, students have just completed their Phase III medicine rotation.

1. Examine the alternatives and indicate those alternatives that a Phase III student, who has just finished his or her medicine rotation should ELIMINATE AS INCORRECT. The basis of your decision is purely judgmental, but it combines your own knowledge of medicine and medical education with an understanding of how students at this level need to be able to use their knowledge to work their way toward clinical decisions.

2. As you examine the alternatives, please tell me what you are thinking so that I can keep track of your line of reasoning. Vocalize the thoughts as soon as possible while you read the item. Take as much time as you need to read each item.

3. If you believe that students at this stage who reason competently should know that option number 4 is incorrect, then you would cross off 4 and explain your reasoning. Perhaps you believe that competent reasoning should tell students that options 1, 3, and 4 are incorrect, but they may not be to the point of distinguishing between 2 and 5, then you would cross off 1, 3, and 4 and give your reasons. In some items you may believe that any student at this stage should get the item correct. For such an item, you would cross off all but the correct alternative. There may also be items that only exceptional students can answer successfully. In your opinion, a competent student may not possess the knowledge and reasoning skills necessary to answer the item. In such cases you would not cross off any alternative.

PART 2

EXAMINE THE ALTERNATIVES AND INDICATE THOSE ALTERNATIVES THAT A PHASE III STUDENT, WHO HAS JUST FINISHED HIS OR HER MEDICINE ROTATION, SHOULD ELIMINATE AS INCORRECT. THE BASIS OF YOUR DECISION IS PURELY JUDGMENTAL, BUT IT COMBINES YOUR OWN KNOWLEDGE OF MEDICINE AND MEDICAL EDUCATION WITH AN UNDERSTANDING OF HOW STUDENTS AT THIS LEVEL NEED TO USE THEIR KNOWLEDGE TO WORK THEIR WAY TOWARD CLINICAL DECISIONS.

PART 3

IN THIS ITEM, USE THE SCALE BELOW TO RATE THE COMPETENCE OF STUDENTS WHO CHOSE EACH DISTRACTER.

LOW HIGH

1 2 3 4

1 IF STUDENTS CHOOSE THIS DISTRACTER, I WOULD HAVE SERIOUS DOUBTS ABOUT THEIR COMPETENCE.

2 IF STUDENTS CHOOSE THIS DISTRACTER, I WOULD HAVE SOME DOUBTS ABOUT THEIR COMPETENCE.

3 STUDENTS COULD CHOOSE THIS DISTRACTER AND STILL BE CONSIDERED COMPETENT.

4 THIS WOULD BE AN ATTRACTIVE DISTRACTER FOR COMPETENTLY THINKING STUDENTS TO CHOOSE.

PART 4

LOOK AT THE CORRECT ANSWER AND THINK ABOUT STUDENTS AT THE PHASE III LEVEL. TO GET THE CORRECT ANSWER, HOW WOULD THEY REASON TO ARRIVE AT THIS ANSWER?

Appendix C: Contingency Tables of the Clinicians' Decisions Comparisons

2 X 2 Contingency Table: Item 317 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	34	7	41
	3, 4	1	2	3
	Total	35	9	44

2 X 2 Contingency Table: Item 771 G1

		Nedelsky Type		
		Elim.	Not Elim.	Tot.
Comp. Rating	1, 2	32	9	41
	3, 4	0	3	3
	Total	32	12	44

2 X 2 Contingency Table: Item 317 G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	28	5	33
	3, 4	1	2	3
	Total	29	9	36

2 X 2 Contingency Table: Item 771 G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	29	8	37
	3, 4	0	3	3
	Total	29	11	40

2 X 2 Contingency Table: Item 317 G1N

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	6	2	8
	3, 4	0	0	0
	Total	6	2	8

2 X 2 Contingency Table: Item 771 G1N

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	2	1	3
	3, 4	0	1	1
	Total	2	2	4

2 X 2 Contingency Table: Item 773 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	20	7	27
	3, 4	0	9	9
	Total	20	16	36

2 X 2 Contingency Table: Item 347 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	21	4	25
	3, 4	3	15	18
	Total	24	19	43

2 X 2 Contingency Table: Item 773 G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	14	4	18
	3, 4	0	6	6
	Total	14	10	24

2 X 2 Contingency Table: Item 347 G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	17	3	20
	3, 4	3	2	15
	Total	20	15	35

2 X 2 Contingency Table: Item 773 G1N's

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	6	3	9
	3, 4	0	3	3
	Total	6	6	12

2 X 2 Contingency Table: Item 347 G1N's

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	4	1	5
	3, 4	0	3	3
	Total	4	4	8

2 X 2 Contingency Table: Item 031 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	32	5	37
	3, 4	0	7	7
	Total	32	12	44

2 X 2 Contingency Table: Item 060 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	28	1	29
	3, 4	2	5	7
	Total	30	6	36

2 X 2 Contingency Table: Item 031 G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	29	5	34
	3, 4	0	6	6
	Total	29	11	40

2 X 2 Contingency Table: Item 060 G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	13	1	14
	3, 4	0	2	2
	Total	13	3	16

2 X 2 Contingency Table:Item 031 G1N's

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	3	0	3
	3, 4	0	1	1
	Total	3	1	4

2 X 2 Contingency Table: Item 060 G1N

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	15	0	15
	3, 4	2	3	5
	Total	17	3	20

2 X 2 Contingency Table: Item 332 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	9	1	10
	3, 4	1	9	10
	Total	10	10	20

2 X 2 Contingency Table: Item 267 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	14	1	15
	3, 4	3	2	5
	Total	17	3	20

2 X 2 Contingency Table: Item 332 G1E's

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	7	1	8
	3, 4	1	3	4
	Total	8	4	12

2 X 2 Contingency Table: Item 267 G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	7	1	8
	3, 4	3	1	4
	Total	10	2	12

2 X 2 Contingency Table:Item 332 G1N's

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	2	0	2
	3, 4	0	6	6
	Total	2	6	8

2 X 2 Contingency Table: Item 267 G1N

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	7	0	7
	3, 4	0	1	1
	Total	7	1	8

2 X 2 Contingency Table: Item 733 G1

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	23	0	23
	3, 4	1	8	9
	Total	24	8	32

2 X 2 Table: Item 582 G1 & G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	31	1	32
	3, 4	3	9	12
	Total	34	10	44

2 X 2 Contingency Table: Item 733, G1E

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	10	0	10
	3, 4	1	1	2
	Total	11	1	12

2 X 2 Contingency Table: Item 733 G1N

		Nedelsky Type		
		Elim.	Not Elim.	Total
Comp. Rating	1, 2	13	0	13
	3, 4	0	7	7
	Total	13	7	20

Appendix D: Distractor Verification Analysis

Distractor Elimination Verification Instructions

A summary of the categories and codes for each distractor prefaces each item. Each item has a spreadsheet, listing the clinicians and students to be coded. First, identify the rationale in the protocol, then code it according to the number on the code sheet. Please fill in the reason for elimination for each option on the spreadsheet.

Appendix E: Clinicians' Perceptions Verification Analysis

Clinicians' Perceptions Verification Analysis Instructions

Each item has a spreadsheet, listing the clinicians and students to be coded. Identify the clinicians' perception of how a student would solve the item in the clinicians' protocol. Read the students' protocols and match the clinicians' explanation of reasoning to the students' protocols. Please fill in the clinicians' perceptions on the spreadsheet.

For Clinicians and Students:

Atomistic: The individual solved the item by using pieces of information, discreet bits of knowledge, facts, definitions. Generated hypotheses from the data in the stem and alternatives. The hypotheses may be incorrect, and many are generated. They try to match the data in the alternatives with the scenario. The item is solve by addressing each alternative individually.

Holistic: The individual generated a hypothesis while reading the stem, before scanning the list of alternatives. They offer a hypothesis before reading the alternatives, usually the keyed response. Views the scenario from a global perspective.

For Clinicians Only:

Expectation stated: Clinicians stated the behaviour/knowledge level they thought a student would need to solve the item. They may say, "I would expect them to know that." They would separate their knowledge level from that of a student by saying, "Students may not have the experience to know this."

Did Not Explain Reasoning Process of Students: Clinicians proceed to explain or describe the disease, and/or treatment without acknowledging how a student would think through the item. They may say, "In practice we treat patients in this way."